Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Diagnosis Everywhere: Lightweight AI That Detects Disease from Scans and Records on Limited Hospital Hardware

Md Sahid Hossain[1], Kallol Chakraborty Shekhor[2], Md Anwar Hossain[3], and Md Abedur Rahman[3]

[1] *Senior Software Engineer, Prime Tech Solutions Ltd., Dhaka, Bangladesh*

[2] *Master of Science in Information Studies, Trine University, Allen Park, MI, U.S.A*

[3] *Master's in Computer Science, Maharishi International University, Fairfield, IA, U.S.A*

| ABSTRACT

Federated learning enables collaborative medical model training without centralizing raw institutional data, yet its practical use remains constrained by non IID distributions, communication overhead, limited computing resources, and the need to support multiple diagnostic tasks across different modalities. This paper presents the proposed model, an efficiency aware federated multi task learning framework for medical imaging and clinical text applications. For imaging tasks, the model uses a frozen SqueezeNet encoder with task specific prediction heads and applies entropy adaptive structured pruning, where each client's pruning rate is adjusted according to local label distribution heterogeneity. For clinical text, BioClinicalBERT is adapted using LoRA modules inserted into the query and value projections, followed by structured top 3 layer pruning to reduce inference and communication cost. Experiments are conducted on fundus based eye disease classification, COVID 19 radiography classification, and five unified clinical text datasets. Under non IID conditions, the proposed model achieves 90.27 ± 0.44% accuracy and 0.903 F1 for eye disease classification, and 78.51 ± 0.53% accuracy and 0.784 F1 for COVID 19 radiography. Compared with FedAvg, FedProx, SCAFFOLD, FedBN, AutoFLIP, and fixed pruning, the adaptive variant provides the strongest imaging performance while reducing communication to 3.14 MB per round and active parameters to 0.87M. In clinical text modeling, FLUTE based LoRA fine tuning achieves 91.76 ± 0.36% accuracy and 0.913 F1. With top 3 layer pruning, the model retains 90.94 ± 0.40% accuracy while reducing parameters to 86M, FLOPs to 79.4G, latency to 79 ms, communication to 0.85 MB per round, and estimated energy demand by 22.6%. These results indicate that adaptive compression and parameter efficient adaptation can improve federated medical learning under heterogeneous and resource constrained conditions.

| KEYWORDS

Federated learning, multitask learning, medical imaging, clinical text diagnosis, adaptive pruning, LoRA

| ARTICLE INFORMATION

ACCEPTED: 01 October 2023 **PUBLISHED:** 07 November 2023 **DOI:** 10.32996/fcsai.2023.2.2.6

1. Introduction

Federated learning has become an important paradigm for medical artificial intelligence because it enables collaborative model training across institutions without centralizing raw patient data. This is particularly relevant in healthcare, where medical images, clinical notes, diagnostic records, and patient histories are governed by strict institutional, ethical, and regulatory constraints. By keeping data within local environments, federated learning offers a practical route for collaborative model development across hospitals, clinics, and research centers. However, real medical federated learning remains difficult because participating institutions often differ in patient populations, disease prevalence, imaging devices, annotation standards, clinical workflows, and computing capacity.One major challenge is statistical heterogeneity. Standard federated optimization assumes that client datasets are broadly aligned, but medical data is rarely uniform across institutions. In eye disease screening, one site may contain more diabetic retinopathy cases, while another may contain more glaucoma or normal cases. In chest radiography, the

distribution of COVID 19, pneumonia, lung opacity, and normal cases may vary substantially across hospitals. This non IID structure can cause local training updates to drift toward institution specific patterns, weakening global convergence and reducing diagnostic reliability. Therefore, federated medical models require mechanisms that preserve local adaptability without sacrificing collaborative learning.A second challenge is the multi task nature of clinical practice. A real healthcare network rarely trains only one diagnostic model. Different institutions may require fundus image classification, chest X ray interpretation, symptom based disease prediction, diabetes risk assessment, and heart failure prediction at the same time. These tasks differ in modality, label space, architecture, and computational demand. A single global model is often too rigid for such diversity, while fully independent local models fail to exploit shared medical knowledge. Federated multi task learning can address this issue by sharing common representations while maintaining task specific adaptation, but existing designs often increase communication and memory burden.Resource limitation is another critical barrier. Medical institutions do not always have high performance computing infrastructure, and repeated communication of large model updates can become impractical. This problem becomes more severe when transformer based clinical language models are used, because full parameter tuning and transmission require substantial memory and bandwidth. Medical imaging models also create communication overhead when full backbones or task branches are exchanged during federated rounds. For federated learning to become practical in clinical environments, model updates must be compact, computation must be controlled, and adaptation should focus only on components that are necessary for each task.

This paper presents the proposed model, an efficiency aware federated multi task learning framework for medical imaging and clinical text applications. The framework is designed around a simple principle. Shared representations should remain stable, while only compact task adaptive components should be trained, compressed, and communicated. For medical imaging, the proposed model uses a frozen SqueezeNet encoder to extract transferable visual features and avoid encoder gradient transmission. Each diagnostic task uses a lightweight prediction head, which is compressed before communication through structured pruning.A key component of the imaging branch is heterogeneity aware adaptive pruning. Instead of applying one fixed pruning rate to all clients, the proposed model adjusts the pruning rate according to the entropy of each client's local label distribution. Clients with more balanced data can tolerate stronger compression, while clients with more skewed distributions retain more predictor capacity. This design connects model compression directly to client heterogeneity and improves the balance between communication efficiency and diagnostic performance under non IID conditions.For clinical text, the proposed model uses BioClinicalBERT as the domain specific language encoder and applies LoRA adapters to the query and value projections. The backbone remains frozen, and only the low rank adapter parameters are trained and communicated during federated optimization. To further reduce computation, structured top 3 layer pruning is applied after adaptation. This removes upper transformer layers while preserving lower layers that retain foundational clinical language representations. The result is a compact federated clinical text model with reduced communication, lower latency, and limited performance loss.

2. Related Work

Federated learning has been widely studied as a collaborative learning paradigm for healthcare because it allows multiple institutions to train shared models while keeping raw data within local environments. Healthcare surveys consistently identify medical imaging, smart healthcare monitoring, and clinical decision support as major application areas, but they also highlight unresolved challenges in non-IID data, communication cost, privacy leakage from model updates, and limited real-world validation [1–3]. These issues are central to medical deployment because hospitals differ in patient populations, scanner protocols, disease prevalence, annotation standards, and computational infrastructure. Existing healthcare federated learning studies establish the need for decentralized medical AI, but they do not fully address the combined requirements of multi-task learning, adaptive compression, and parameter-efficient clinical text adaptation.Non-IID data is one of the main causes of unstable federated optimization. FedProx extends FedAvg by adding a proximal term to the local objective, reducing client drift under statistical and systems heterogeneity [4]. SCAFFOLD addresses the same issue through control variates that correct biased local updates caused by heterogeneous client distributions [5]. FedBN keeps batch normalization statistics local to handle feature shift, which is especially relevant for medical imaging where scanner or acquisition differences can alter the feature distribution across institutions [6]. These methods improve optimization stability, but they mainly focus on global convergence and do not directly design adaptive model compression based on local medical heterogeneity.

Federated multi-task learning is important for healthcare because institutions often require related but non-identical diagnostic models. FedBone proposes a large-scale federated multi-task framework that separates general and task-specific components while addressing gradient conflict among heterogeneous tasks [7]. FedMRL extends heterogeneous federated learning through Matryoshka representation learning, allowing clients with different model structures to exchange multi-granular representations [8]. LeTS reduces computation and storage burden through computation and parameter sharing across multiple tasks [9]. These studies support the value of shared representations and task-specific adaptation, but they do not provide a unified medical framework that combines imaging-specific adaptive pruning with clinical text parameter-efficient fine-tuning.Communication

cost remains a major barrier in federated learning because repeated transmission of full model parameters becomes expensive in distributed clinical environments. AutoFLIP uses federated loss exploration to drive adaptive hybrid pruning and reduce communication and computation under non-IID settings [10]. SparseGPT shows that large transformer models can be pruned in one shot with limited degradation, demonstrating the potential of structured sparsity for large-scale language models [11]. FedSpaLLM further extends pruning to federated large language models by allowing clients to prune locally while maintaining communication efficiency [12]. These works motivate model compression, but they do not adapt pruning intensity using medical client label entropy, nor do they integrate pruning with federated multi-task imaging and clinical language modeling in a single framework.Transformer-based clinical language modeling is difficult in federated settings because full fine-tuning requires large parameter updates and high memory use. LoRA reduces this burden by freezing pretrained weights and injecting trainable low-rank matrices into transformer layers, which greatly reduces the number of trainable parameters [13]. FedPrompt applies prompt tuning in federated learning and reduces communication by transmitting soft prompt parameters rather than full language model updates [14]. FedBPT further studies black-box prompt tuning for federated large language models, where clients do not need direct access to model parameters [15]. These approaches improve communication efficiency, but prompt-based methods can be less expressive for complex clinical diagnosis tasks, while LoRA-based adaptation still requires further compression for resource-constrained deployment.

Practical federated experiments also depend on stable simulation and deployment frameworks. Flower provides a flexible federated learning framework designed to support heterogeneous client environments and scalable FL experimentation [16]. FLUTE offers a scalable simulation environment for high-performance federated learning workloads and is useful for transformer-based federated experiments where memory handling and tokenizer management affect convergence stability [17]. These frameworks support reproducible federated evaluation, but framework choice alone does not solve the algorithmic challenges of non-IID learning, adaptive compression, and parameter-efficient model adaptation.The existing literature leaves three important gaps. First, healthcare federated learning studies commonly focus on single-task diagnosis or broad privacy-aware collaboration, while practical clinical systems require multiple diagnostic tasks across imaging and text. Second, non-IID optimization methods improve convergence but do not adjust compression intensity according to local medical data heterogeneity. Third, pruning and parameter-efficient tuning are often studied separately, although resource-constrained clinical deployment requires both compact communication and modality-specific adaptation. The proposed model addresses these gaps by combining a frozen imaging encoder with entropy-adaptive pruned task heads, and a frozen clinical language backbone with LoRA-based adaptation and structured top-layer pruning.

3.1 Dataset Description

The proposed model was evaluated using six datasets covering two medical modalities: medical imaging and clinical text. The imaging component includes fundus based eye disease classification and chest X ray based pulmonary disease classification, while the clinical text component includes diabetes prediction, symptom based disease prediction, disease profile classification, and heart failure prediction. This dataset selection was designed to test the proposed model under different input types, label spaces, task complexities, and federated client settings. The full dataset specification is summarized in Table 1.The Eye Diseases Classification dataset contains 4,217 RGB fundus images resized to 224 × 224 pixels. It consists of four classes: cataract, diabetic retinopathy, glaucoma, and normal. All samples were retained after cleaning, resulting in 4,217 usable images. The dataset was divided into training, validation, and test subsets using a 70/15/15 split and distributed across three federated clients. This dataset was used to evaluate the ability of the proposed model to learn transferable retinal representations under federated medical imaging conditions.

The COVID 19 Radiography dataset contains 21,165 chest X ray images resized to 224 × 224 pixels. It includes four diagnostic categories: COVID 19, pneumonia, lung opacity, and normal. No samples were removed during cleaning, so the cleaned dataset retained all 21,165 images. The dataset was split into training, validation, and test subsets using a 70/15/15 ratio and assigned to three federated clients. This task provides a more challenging imaging setting because chest X ray abnormalities can show overlapping radiographic patterns across disease classes.For the clinical text branch, two diabetes prediction datasets were combined into a single binary classification task. The combined dataset originally contained 9,538 samples, of which 8,964 remained after cleaning. The cleaned records were split using an 80/10/10 training, validation, and test ratio and distributed across two clients. This dataset was used to assess binary disease prediction from structured clinical text representations.The Symptom2Disease dataset contains 1,200 symptom based clinical text samples mapped to 24 disease classes. After cleaning, 1,184 records were retained. The dataset was split into training, validation, and test subsets using an 80/10/10 ratio and assigned to one federated client. This dataset evaluates the model's ability to distinguish multiple diseases from symptom descriptions.The Disease Symptoms and Patient Profiles dataset contains 1,317 clinical text records across 25 disease categories. All samples were retained after cleaning. The dataset was divided using an 80/10/10 split and assigned to one client. This dataset increases the diversity of clinical text labels and supports evaluation under a broader multi class disease classification setting.The

Heart Failure Prediction dataset contains 918 samples for binary clinical prediction. After cleaning, 906 samples were retained. The dataset was divided into training, validation, and test subsets using an 80/10/10 split and distributed across two federated clients. This dataset adds an additional binary diagnostic task to the clinical text branch.

Table 1. Dataset specification used in the Proposed Model experiments.

Dataset	Modality	Classes	Total	Cleaned	Split (Tr/Val/Te)	Clients
Eye Diseases Classification	Fundus RGB 224 x 224	4 (Cataract, DR, Glaucoma, Normal)	4,217	4,217	70/15/15	3
COVID-19 Radiography	CXR 224 x 224	4 (COVID-19, Pneumonia, Lung Opacity, Normal)	21,165	21,165	70/15/15	3
Diabetes Prediction (two combined datasets)	Clinical text	2	9,538	8,964	80/10/10	2
Symptom2Disease	Clinical text	24	1,200	1,184	80/10/10	1
Disease Symptoms and Patient Profiles	Clinical text	25	1,317	1,317	80/10/10	1
Heart Failure Prediction	Clinical text	2	918	906	80/10/10	2

3.2 Data Preprocessing

All datasets were preprocessed before federated training to ensure consistent input formatting, remove invalid records, and reduce modality specific noise. Because the proposed model operates on both medical images and clinical text, preprocessing was performed separately for the imaging and text branches.For the medical imaging branch, all fundus and chest X ray images were first checked for unreadable files, corrupted samples, and invalid labels. No samples were removed from the Eye Diseases Classification and COVID 19 Radiography datasets after this screening. Each image was resized to 224 × 224 pixels to match the input requirement of the SqueezeNet based visual encoder. RGB fundus images were retained in three channels, while chest X ray images were standardized to the same input format to ensure compatibility with the shared imaging pipeline.Pixel intensities were normalized to a fixed numerical range before model training. This step reduces scale variation across images and stabilizes gradient updates during local client optimization. For the imaging datasets, standard augmentation was applied only to the training split. The augmentation process included random rotation, horizontal flipping, and mild cropping or resizing operations. Validation and test images were not augmented, ensuring that evaluation reflected the original data distribution.

For the clinical text branch, all tabular and symptom based records were converted into a unified text label format. Records with missing disease labels, invalid entries, duplicate noise, or non informative clinical fields were removed during cleaning. The two diabetes prediction datasets were combined into a single binary classification task after standardizing their fields. For each clinical text dataset, relevant attributes such as symptoms, risk indicators, patient profile information, or diagnostic descriptors were concatenated into a single textual input. Labels were then mapped into task specific class indices.Text preprocessing included lower level cleaning, removal of empty records, standardization of class names, and conversion of structured clinical fields into natural language style inputs. The cleaned text was tokenized using the BioClinicalBERT tokenizer. A maximum sequence length of 512 tokens was used. Inputs longer than this limit were truncated, while shorter inputs were padded to a fixed length. Attention masks were generated to distinguish real tokens from padding tokens during transformer encoding.After preprocessing, each dataset was divided into training, validation, and test subsets according to its predefined split. Imaging datasets used a 70/15/15 split, while clinical text datasets used an 80/10/10 split. The training subsets were distributed across

federated clients according to the experimental design. For non IID experiments, client partitions were created to introduce label distribution skew while keeping validation and test evaluation separate from local training. All preprocessing steps were applied before federated training, and no test samples were used during model fitting, pruning decisions, LoRA adaptation, or threshold selection.

3.3 Proposed Model Architecture

3.3.1 Overall Federated Multi Modal Design

The proposed model is designed as an efficiency aware federated multi task framework for medical imaging and clinical text learning. The architecture contains two modality specific branches that operate under a shared federated training protocol. The first branch processes medical images, including fundus images and chest X ray images, while the second branch processes clinical text records derived from symptom descriptions, disease profiles, and structured diagnostic attributes. Each branch uses a frozen domain specific encoder to extract stable representations and trains only compact task adaptive components. This design reduces communication overhead, limits local computation, and supports heterogeneous diagnostic tasks without centralizing raw medical data.At the client side, each institution performs local training using its own private data. Imaging clients train lightweight task specific prediction heads on top of a shared visual encoder, while clinical text clients train low rank adapter modules on top of a frozen clinical language model. Before communication, client updates are compressed through modality appropriate efficiency mechanisms. Imaging updates are reduced using entropy adaptive structured pruning, whereas clinical text updates are restricted to LoRA adapter parameters and further compressed through structured top layer pruning. The server receives only these compact trainable components, aligns them when necessary, performs weighted aggregation, and redistributes the updated global components for the next communication round.

3.3.2 Medical Imaging Branch with Adaptive Pruned Task Heads

The imaging branch uses SqueezeNet as a compact visual encoder because it provides a favorable balance between representation capacity and computational efficiency. The encoder is kept frozen during federated optimization to preserve general visual features and avoid transmitting encoder gradients. This is important in resource constrained medical settings because full backbone communication increases bandwidth cost and may amplify instability under non IID data distributions. The frozen encoder converts each fundus or chest X ray image into a compact feature representation that is passed to a task specific prediction head.Each diagnostic task uses its own trainable head. For fundus imaging, the head predicts cataract, diabetic retinopathy, glaucoma, and normal classes. For chest radiography, the head predicts COVID 19, pneumonia, lung opacity, and normal classes. This separation allows the model to support multiple medical imaging tasks without forcing all clients into a single output space. It also limits task interference because each head specializes in one diagnostic objective while still benefiting from a shared visual representation.To improve communication efficiency, each client applies structured pruning to its local prediction head before sending updates to the server. The pruning process removes low importance filters or units from the task head rather than pruning individual weights in an unstructured manner. This preserves hardware friendly model structure and reduces the number of transmitted parameters. The key architectural feature is that pruning is adaptive rather than fixed. The pruning intensity is determined from the local label distribution. Clients with more balanced label distributions can tolerate stronger pruning, while clients with stronger label skew retain more prediction head capacity. This makes compression sensitive to medical heterogeneity and prevents over pruning on difficult or imbalanced client data.

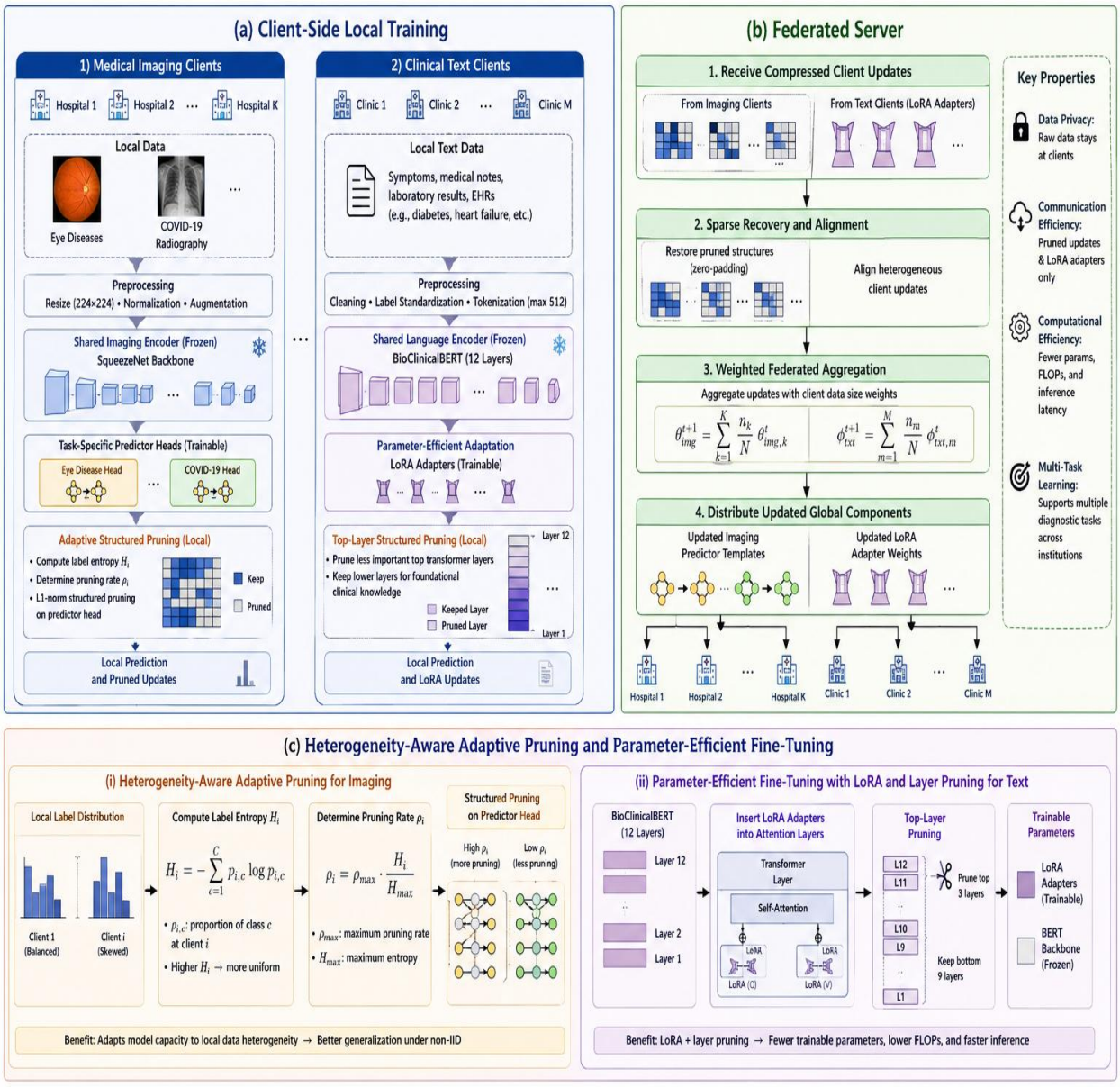

Figure 1. Proposed Model Architecture for Federated Medical Imaging and Clinical Text Learning

3.3.3 Clinical Text Branch with LoRA Adaptation and Top Layer Pruning

The clinical text branch uses BioClinicalBERT as the domain specific language encoder. This choice is suitable for medical text because the model already contains clinical terminology and contextual representations learned from biomedical and clinical corpora. Instead of fully fine tuning all transformer parameters, the proposed model freezes the BioClinicalBERT backbone and inserts LoRA adapters into the query and value projections of the attention layers. Only these adapter parameters are updated during federated training.This design substantially reduces trainable and transmitted parameters while preserving the clinical knowledge stored in the pretrained encoder. It also improves federated feasibility because clients do not need to update or communicate the full transformer backbone. The clinical text inputs are tokenized and passed through the frozen encoder, while the LoRA adapters provide task specific adaptation for disease prediction. This enables the model to learn from heterogeneous clinical text datasets while keeping communication cost low.After LoRA based adaptation, structured top layer pruning is applied to reduce the transformer depth. The architecture removes the upper transformer layers while retaining the lower layers that encode foundational clinical language patterns. This pruning strategy is more suitable than alternating layer removal because disrupting intermediate layers can break the hierarchical flow of contextual representations. Top layer pruning therefore provides a cleaner compression path, lowering parameters, FLOPs, latency, and communication cost while preserving most of the diagnostic performance.

3.3.4 Server Side Aggregation and Final Prediction Workflow

The federated server coordinates the learning process without accessing raw medical images or clinical text. During each communication round, imaging clients transmit compressed task head updates, while text clients transmit LoRA adapter updates. For the imaging branch, the server first aligns pruned prediction heads so that heterogeneous client structures can be aggregated consistently. Removed positions are treated as inactive components during aggregation, while retained components contribute according to client data size. For the clinical text branch, the server aggregates adapter parameters rather than full BioClinicalBERT weights.After aggregation, the server redistributes the updated global task heads and adapter parameters to participating clients. Each client then performs the next round of local training using its own private data. At inference time, medical images pass through the frozen visual encoder and the relevant task specific head, while clinical text inputs pass through BioClinicalBERT with the learned LoRA adapters and the pruned transformer stack. This workflow allows the proposed model to support multiple diagnostic tasks across imaging and text modalities while maintaining compact communication and controlled computation.

3.4 Experimental Configuration

3.4.1 Federated Client Setup

The proposed model was evaluated under a simulated federated learning environment with separate client groups for medical imaging and clinical text tasks. The imaging branch used six clients in total, with three clients assigned to the Eye Diseases Classification task and three clients assigned to the COVID 19 Radiography task. This setting reflects a multi task medical federation in which different institutions may participate in different but related diagnostic objectives.For the clinical text branch, the cleaned text datasets were distributed across six clients. The two combined diabetes prediction datasets were assigned to two clients, the Symptom2Disease dataset to one client, the Disease Symptoms and Patient Profiles dataset to one client, and the Heart Failure Prediction dataset to two clients. This client allocation was used to evaluate whether the proposed model can handle unequal label spaces, different dataset sizes, and heterogeneous clinical text distributions within the same federated protocol.

3.4.2 IID and Non IID Partitioning

Two data distribution settings were used. The IID setting served as the controlled baseline, where class distributions were kept approximately balanced across clients. The non IID setting was used to simulate realistic institutional heterogeneity. For the imaging experiments, non IID partitions were generated with label distribution skew using Dirichlet based partitioning with $\alpha = 0.5$. This created uneven class proportions across clients while preserving the same diagnostic label space within each task.The non IID setting is important because medical institutions rarely observe identical disease distributions. Some hospitals may contain more normal cases, while others may contain more disease specific cases. The non IID experiment therefore tests whether the proposed model can maintain stable performance when local client updates are influenced by skewed class distributions.

3.4.3 Imaging Model Configuration

The imaging branch used SqueezeNet as the frozen visual encoder. All input images were resized to 224×224 pixels and normalized before training. The encoder was kept frozen throughout the federated rounds to reduce communication cost and preserve stable visual representations. Only the task specific prediction heads were trained locally and transmitted to the server.Each imaging client trained its local prediction head using the assigned task data. Before communication, the prediction head was compressed using structured pruning. Two pruning variants were evaluated. The fixed pruning variant used a constant pruning rate of $\rho = 0.25$ for all clients. The adaptive pruning variant adjusted the pruning rate according to local label distribution heterogeneity. This comparison was used to determine whether client aware pruning provides better accuracy efficiency trade off than uniform compression.

3.4.4 Clinical Text Model Configuration

The clinical text branch used BioClinicalBERT as the base language encoder. The transformer backbone was frozen, and only LoRA adapters were trained. LoRA adapters were inserted into the query and value projection layers, with rank $r = 8$. This configuration limits trainable parameters while preserving the domain specific representation capacity of BioClinicalBERT.Federated text training was conducted using FLUTE because it provided more stable memory handling than Flower in the controlled comparison. The final text model used six clients and 20 communication rounds. After LoRA based federated fine tuning, structured transformer pruning was evaluated. Two pruning strategies were compared: alternating layer pruning and top 3 layer pruning. Top 3 layer pruning was selected as the stronger configuration because it preserved more diagnostic performance while reducing parameters, FLOPs, latency, and communication cost.

3.4.5 Baselines and Evaluation Metrics

The imaging branch was compared with centralized training, FedAvg, FedProx, SCAFFOLD, FedBN, AutoFLIP, the proposed model with fixed pruning, and the proposed model with adaptive pruning. These baselines were selected to cover standard aggregation, drift correction, feature normalization personalization, pruning based compression, and centralized upper bound performance.The clinical text branch was evaluated using head only federated tuning, LoRA based federated tuning, LoRA without pruning, LoRA with alternating layer pruning, and LoRA with top 3 layer pruning. Flower and FLUTE were also compared under controlled conditions to identify the more stable federated training framework for transformer based clinical text modeling.Performance was measured using accuracy, precision, recall, F1 score, and macro AUC where applicable. Efficiency was evaluated using communication cost per round, active parameters, FLOPs, inference latency, training time per round, and estimated energy reduction. Reported values include mean and standard deviation where repeated runs were available, allowing performance stability to be assessed across model variants.

3.4.6 Implementation Environment

All experiments were implemented in Python using PyTorch based deep learning pipelines. The clinical text branch used the Hugging Face Transformers and PEFT libraries for BioClinicalBERT and LoRA adaptation. Flower and FLUTE were used for federated training comparisons. Experiments were conducted on GPU based hardware, with latency measured under controlled inference settings. The same preprocessing, split policy, and evaluation protocol were maintained across compared methods to support fair experimental comparison.

4. Results and Analysis

4.1 Eye Diseases Classification

The proposed model achieved strong performance on the Eye Diseases Classification task under both IID and non IID settings. Under IID partitioning, the model obtained 92.84 ± 0.31% accuracy, with precision of 0.9291, recall of 0.9284, F1 score of 0.9279, and macro AUC of 0.969. Under non IID partitioning with $\alpha = 0.5$, accuracy decreased to 90.27 ± 0.44%, while F1 score remained high at 0.9031 and macro AUC remained 0.953. The drop of 2.57 percentage points indicates that the model retained stable retinal feature representation despite client level label skew. The number of rounds required for convergence increased from 24 under IID conditions to 32 under non IID conditions, showing that heterogeneity slowed optimization but did not cause severe performance collapse. These results are reported in Table 2.

Table 2. Eye Diseases classification performance of Proposed Model .

Setup	Accuracy (%)	Precision	Recall	F1-score	Macro-AUC	Rounds to converge
IID	92.84 ± 0.31	0.9291	0.9284	0.9279	0.969	24
Non-IID ($\alpha = 0.5$)	90.27 ± 0.44	0.9046	0.9027	0.9031	0.953	32

4.2 COVID 19 Radiography Classification

The COVID 19 Radiography task showed lower overall performance and greater sensitivity to non IID partitioning. Under IID conditions, the proposed model achieved 83.62 ± 0.35% accuracy, with precision of 0.8378, recall of 0.8362, F1 score of 0.8358, and macro AUC of 0.921. Under non IID conditions, accuracy declined to 78.51 ± 0.53%, with F1 score of 0.7847 and macro AUC of 0.894. The performance drop of 5.11 percentage points is larger than that observed for the eye disease task, suggesting that chest X ray classification is more affected by client distribution skew and class overlap. The model also required 38 rounds to converge in the non IID setting compared with 29 rounds in the IID setting. These findings indicate that pulmonary image classification remains the more difficult imaging task in the proposed federated setting, as shown in Table 3.

Table 3. COVID-19 Radiography classification performance of Proposed Model .

Setup	Accuracy (%)	Precision	Recall	F1-score	Macro-AUC	Rounds to converge
IID	83.62 ± 0.35	0.8378	0.8362	0.8358	0.921	29
Non-IID (α = 0.5)	78.51 ± 0.53	0.7886	0.7851	0.7847	0.894	38

4.3 Comparison with Federated Imaging Baselines

Under non IID conditions, the adaptive version of the proposed model outperformed all federated imaging baselines while using fewer active parameters and lower communication cost. For eye disease classification, the proposed adaptive model achieved 90.27 ± 0.44% accuracy and 0.903 F1, exceeding FedAvg, FedProx, SCAFFOLD, FedBN, AutoFLIP, and the fixed pruning variant. Compared with FedAvg, it improved eye disease accuracy by 2.33 percentage points and F1 score by 0.026. Compared with the fixed pruning variant, adaptive pruning improved eye disease accuracy by 0.84 percentage points, indicating that client aware pruning is more effective than using a uniform pruning rate.For COVID 19 radiography, the same adaptive model achieved 78.51 ± 0.53% accuracy and 0.784 F1, again outperforming all federated baselines. The improvement over FedAvg was 3.59 percentage points, while the improvement over the fixed pruning variant was 1.30 percentage points. These gains are important because the COVID 19 task showed greater distributional sensitivity, suggesting that adaptive pruning helps preserve useful client capacity under harder non IID conditions.The efficiency results also favor the adaptive proposed model. Communication cost was reduced to 3.14 MB per round, compared with 4.74 MB for FedAvg and FedProx, 9.48 MB for SCAFFOLD, and 3.39 MB for the fixed pruning variant. Active parameters decreased to 0.87M, compared with 1.24M for FedAvg, FedProx, SCAFFOLD, and FedBN. Latency was also reduced to 91 ms, compared with 100 ms for standard baselines. Overall, Table 4 shows that the proposed adaptive pruning strategy improves both predictive performance and resource efficiency, rather than improving one at the cost of the other.

Table 4. Comparison of Proposed Model with federated learning baselines under non-IID conditions.

Method	Eye Accuracy (%)	Eye F1	COVID-19 Accuracy (%)	COVID-19 F1	Communication (MB/round)	Latency (ms)	Active parameters
Centralized upper bound	94.73 ± 0.26	0.946	86.41 ± 0.32	0.862	—	100	1.24M
FedAvg [1]	87.94 ± 0.69	0.877	74.92 ± 0.88	0.745	4.74	100	1.24M
FedProx (μ = 0.01) [54]	88.52 ± 0.62	0.882	75.74 ± 0.81	0.753	4.74	100	1.24M
SCAFFOLD [55]	89.16 ± 0.55	0.889	76.48 ± 0.74	0.761	9.48	103	1.24M
FedBN [19]	88.61 ± 0.63	0.883	75.93 ± 0.77	0.756	4.71	100	1.24M
AutoFLIP [30]	86.78 ± 0.81	0.864	74.06 ± 0.97	0.737	3.51	93	0.93M
Proposed Model (fixed ρ = 0.25)	89.43 ± 0.56	0.892	77.21 ± 0.70	0.769	3.39	94	0.91M
Proposed Model (adaptive ρ, ours)	**90.27 ± 0.44**	**0.903**	**78.51 ± 0.53**	**0.784**	**3.14**	**91**	**0.87M**

4.4 Federated Clinical Text Framework Comparison

The clinical text experiments first compared Flower and FLUTE under controlled federated settings. With 20 rounds and six clients, Flower achieved 85.37 ± 0.72% accuracy, validation loss of 2.27, training time of 314 seconds per round, and three out of memory events. Under the same controlled setting, FLUTE achieved 89.18 ± 0.49% accuracy, validation loss of 1.83, training time of 248 seconds per round, and no out of memory events. This corresponds to a 3.81 percentage point accuracy gain and a substantial improvement in memory stability.The FLUTE round ablation further shows the effect of communication rounds on convergence. With only 3 rounds, FLUTE achieved 64.92 ± 1.08% accuracy, indicating under training. Accuracy increased to 87.41 ± 0.58% after 10 rounds and reached 91.76 ± 0.36% in the final 20 round setting. Validation loss decreased consistently from 3.21 to 2.04 and then to 1.52, while training time per round remained stable at approximately 243 to 247 seconds. These results support the use of FLUTE for the final clinical text experiments because it produced higher accuracy, lower validation loss, faster per round training, and greater memory stability. The comparison is summarized in Table 5.

Table 5. Flower and FLUTE comparison with FLUTE round ablation.

Framework	Rounds	Clients	Accuracy (%)	Validation loss	Training time per round (s)	OOM events
Flower	20	6	85.37 ± 0.72	2.27	314	3
FLUTE	20	6 (controlled)	89.18 ± 0.49	1.83	248	0
FLUTE	3	6	64.92 ± 1.08	3.21	246	0
FLUTE	10	6	87.41 ± 0.58	2.04	247	0
FLUTE (final)	**20**	**6**	**91.76 ± 0.36**	**1.52**	**243**	**0**

4.5 Clinical Language Model Variant Analysis

The language model comparison demonstrates the importance of parameter efficient adaptation. The federated head only model without LoRA achieved 76.83 ± 0.76% accuracy and 0.764 F1, which confirms that updating only the classification head is insufficient for heterogeneous clinical text learning. Adding LoRA to the query and value projections improved accuracy to 89.42 ± 0.45% and F1 score to 0.891, showing that low rank adaptation provides substantial task specific learning while keeping communication low.The strongest unpruned federated text model was FL + LoRA Q,V with no pruning, which achieved 91.76 ± 0.36% accuracy, 0.913 F1, precision of 0.916, and recall of 0.914. This result remains below the centralized LoRA upper bound of 94.38 ± 0.24%, but the gap is expected because the federated setting introduces client heterogeneity and restricted parameter sharing. Importantly, the federated LoRA model used only 1.13 MB per round, compared with 412.3 MB per round for the head only setting that retains full communication burden.Structured pruning produced a clear difference between pruning strategies. Alternating layer pruning reduced parameters to 86M and FLOPs to 79.4G, but accuracy dropped sharply to 81.65 ± 0.63% and F1 score to 0.813. In contrast, top 3 layer pruning retained 90.94 ± 0.40% accuracy and 0.906 F1, with precision of 0.908 and recall of 0.907. Compared with the unpruned LoRA model, top 3 pruning caused only a 0.82 percentage point accuracy reduction while reducing latency from 100 ms to 79 ms, communication from 1.13 MB to 0.85 MB per round, and estimated energy demand by 22.6%. These results show that top layer pruning is substantially more suitable than alternating pruning for preserving clinical language representations. The full variant comparison is given in Table 6.

Table 6. Comparative performance of language model variants.

Model	Accuracy (%)	F1-score	Precision	Recall	Parameters	FLOPs	Latency (ms)	Communication (MB/round)	Estimated energy reduction (%)
Centralized LoRA upper bound	94.38 ± 0.24	0.943	0.945	0.944	108M	102.4G	100	—	—

FL head-only without LoRA	76.83 ± 0.76	0.764	0.769	0.767	108M	102.4G	100	412.3	Baseline
FL + LoRA-Q,V (r = 8)	89.42 ± 0.45	0.891	0.894	0.892	108M	102.4G	100	1.13	0.0
FL + LoRA-Q,V (r = 8), no pruning (OM-FT)	91.76 ± 0.36	0.913	0.916	0.914	108M	102.4G	100	1.13	0.0
FL + LoRA-Q,V with alternating-layer pruning (6 layers)	81.65 ± 0.63	0.813	0.817	0.815	86M	79.4G	81	0.91	19.8
FL + LoRA-Q,V with top-3 layer pruning (OM-FT&P)	90.94 ± 0.40	0.906	0.908	0.907	86M	79.4G	79	0.85	22.6

Across imaging and clinical text tasks, the results show that the proposed model provides a consistent accuracy efficiency trade off under federated constraints. In medical imaging, adaptive pruning improves non IID performance while reducing communication and active parameters. In clinical text, LoRA based adaptation substantially improves accuracy over head only tuning, and top 3 layer pruning reduces model cost with limited degradation. The results also show that task difficulty matters. Eye disease classification remains relatively stable under non IID conditions, whereas COVID 19 radiography is more sensitive to client skew and visual class overlap. Overall, the proposed model is most effective when compression is not applied uniformly, but is matched to the modality, task structure, and client heterogeneity.

5. Limitations and Future Work

Although the proposed model improves the accuracy efficiency trade off in federated medical imaging and clinical text learning, several limitations remain. First, the experiments are conducted in a simulated federated environment rather than across real hospitals or clinical networks. The non IID partitions approximate institutional heterogeneity through label distribution skew, but they cannot fully capture real differences in scanner type, imaging protocol, patient demographics, clinical documentation style, annotation quality, or local infrastructure. Therefore, the results should be interpreted as controlled federated evidence rather than direct clinical deployment validation.Second, the current framework provides data localization but does not implement formal privacy protection. Raw medical images and clinical text are not transmitted to the server, but model updates may still leak sensitive information under gradient inversion, membership inference, or malicious server assumptions. For this reason, the proposed model should not be described as fully privacy preserving. Future work should integrate secure aggregation, differential privacy, or trusted execution mechanisms and quantify the effect of these protections on accuracy, convergence, communication, and latency.

Third, the imaging experiments are limited to fundus and chest X ray classification. Although these tasks represent two clinically relevant imaging modalities, they do not cover more complex medical imaging settings such as CT, MRI, ultrasound, histopathology, or multi label radiology reporting. The COVID 19 radiography task also shows a larger non IID performance drop, indicating that the proposed model is still sensitive to subtle inter class overlap and skewed pulmonary feature distributions. Future studies should include broader imaging datasets, external site evaluation, class wise error analysis, calibration assessment, and robustness testing under acquisition shift.Fourth, the clinical text branch uses unified public datasets rather than real hospital notes. Public clinical text datasets are useful for reproducible benchmarking, but they may not fully represent noisy electronic health records, abbreviations, missing fields, temporal patient history, or institution specific documentation practices. Future work should evaluate the model on de identified real world clinical notes and test whether LoRA based adaptation remains stable under larger label spaces, longer documents, and longitudinal patient records.Fifth, efficiency is evaluated through communication cost, active parameters, FLOPs, latency, and estimated energy reduction, but actual energy consumption is not directly measured on deployment hardware. The reported energy reduction is therefore a proxy rather than a

hardware verified measurement. Future experiments should measure power usage, memory footprint, throughput, and thermal behavior on realistic clinical edge servers, hospital workstations, and low resource devices.

6. Conclusion

This study presented the proposed model as an efficiency aware federated multi task framework for medical imaging and clinical text learning under heterogeneous and resource constrained conditions. By combining frozen domain specific encoders, entropy adaptive structured pruning for imaging heads, LoRA based clinical text adaptation, and structured top layer pruning, the model jointly addresses communication cost, client heterogeneity, and task specific adaptation without centralizing raw medical data. The imaging results showed that adaptive pruning improved non IID performance over standard federated baselines while reducing communication and active parameters, with stronger robustness on eye disease classification and a more challenging but still competitive outcome on COVID 19 radiography. The clinical text results further demonstrated that LoRA adaptation substantially improves federated language modeling, while top layer pruning preserves most diagnostic performance with lower latency, communication, FLOPs, and estimated energy demand. These findings indicate that federated medical learning becomes more practical when compression and adaptation are designed as part of the model architecture rather than applied as separate post hoc steps. The proposed model is not a clinically deployed system and still requires validation across real institutions, formal privacy mechanisms, broader imaging and clinical text datasets, and measured hardware efficiency; nevertheless, it provides a reproducible foundation for scalable, resource conscious federated medical AI.

Author Contributions: Md Sahid Hossain led the study, conceptualized the main research direction, designed the proposed efficiency-aware federated multitask learning framework, developed the imaging and clinical text learning strategy, and took primary responsibility for manuscript preparation. Kallol Chakraborty Shekhor contributed to federated learning design, baseline model development, experimental analysis, performance comparison, and manuscript writing. Md Anwar Hossain contributed to dataset preprocessing, coding, implementation of imaging experiments, pruning analysis, evaluation metrics, and result interpretation. Md Abedur Rahman contributed to clinical text modeling, LoRA-based BioClinicalBERT adaptation, computational efficiency analysis, literature review, and manuscript editing. All authors contributed to coding, experimental validation, analysis of results, manuscript revision, and approval of the final version.

Data Availability:The datasets used in this study are publicly available from open-access repositories. The Eye Diseases Classification dataset, COVID-19 Radiography dataset, Diabetes Prediction datasets, Symptom2Disease dataset, Disease Symptoms and Patient Profiles dataset, and Heart Failure Prediction dataset were obtained from publicly accessible Kaggle sources. All datasets were used only after preprocessing, cleaning, and task-specific formatting as described in the methodology section. No private hospital records, identifiable patient information, or institution-owned clinical data were used in this study. The processed dataset splits and experimental configuration can be made available by the corresponding author upon reasonable request, subject to the original dataset licenses and repository access conditions.

Declarations

Clinical Trial Number: Not applicable.

Human Ethics and Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Consent to Participate: Not applicable.

Ethics declaration: Not applicable.

Funding:This research received no external funding.

References

[1] Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. 2021. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5, 1–19.

[2] Nguyen, D. C., Pham, Q. V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O. A., and Hwang, W. J. 2022. Federated learning for smart healthcare: A survey. ACM Computing Surveys, 55(3), 1–37.

[3] Huang, Shih-Cheng, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." *NPJ digital medicine* 3, no. 1 (2020): 136.

[4] Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. 2020. Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.

[5] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. Proceedings of the 37th International Conference on Machine Learning, 5132–5143.

[6] Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. 2021. FedBN: Federated learning on non-IID features via local batch normalization. International Conference on Learning Representations.

[7] Agrawal, Shweta, and Sanjiv Kumar Jain. "Medical text and image processing: applications, issues and challenges." In *Machine learning with health care perspective: machine learning and healthcare*, pp. 237-262. Cham: Springer International Publishing, 2020.

[8] Moon, Jong Hak, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. "Multi-modal understanding and generation for medical images and text via vision-language pre-training." *IEEE Journal of Biomedical and Health Informatics* 26, no. 12 (2022): 6070-6080.

[9] Fu, C., Huang, H., Chen, X., Tian, Y., and Zhao, J. 2021. Learn-to-Share: A hardware-friendly transfer learning framework exploiting computation and parameter sharing. Proceedings of the 38th International Conference on Machine Learning, 3469–3479.

[10] Lobantsev, A. A., N. F. Gusarova, Aleksandra Sergeevna Vatian, Andrey Andreevich Kapitonov, and Anatoly Abramovich Shalyto. "Comparative assessment of text-image fusion models for medical diagnostics." *Информационно-управляющие системы* 5 (108) (2020): 70-79.

[11] Frantar, E., and Alistarh, D. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. Proceedings of the 40th International Conference on Machine Learning.

[12]Hussain, Shah, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A. Sultan. "Modern diagnostic imaging technique applications and risk factors in the medical field: a review." *BioMed research international* 2022, no. 1 (2022): 5164970.

[13] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. 2022. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations.

[14] Zhao, H., Du, W., Li, F., Li, P., and Liu, G. 2022. FedPrompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. arXiv preprint arXiv:2208.12268.
[15] Spasic, Irena, and Goran Nenadic. "Clinical text data in machine learning: systematic review." *JMIR medical informatics* 8, no. 3 (2020): e17984.

[16] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., and Lane, N. D. 2020. Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390.

[17] Garcia, M. H., Manoel, A., Madrigal Diaz, D., Mireshghallah, F., Sim, R., and Dimitriadis, D. 2022. FLUTE: A scalable, extensible framework for high-performance federated learning simulations. arXiv preprint arXiv:2203.13789.