

A Privacy-Preserving Framework for Cross-Institutional Medical Image Analysis Using Vision-Language Models

Yang Jiafeng[1(✉)], Zhukova Natalia Alexandrovna[2], and Stankova Elena Nikolaevna[3]

[1] ITMO University, Saint Petersburg, Russia
tszyafengyang@gmail.com
[2] Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Saint Petersburg, Russia
nazhukova@mail.ru
[3] Saint Petersburg State University, Saint Petersburg, Russia
e.stankova@spbu.ru

Abstract. The healthcare industry faces significant challenges in leveraging patient data across institutions while maintaining privacy, particularly when third-party organizations like insurance companies and banks require medical information for risk assessment. The rapid advancement of large-scale multimodal models, such as Contrastive Language-Image Pre-training (CLIP), holds immense potential for medical applications by enabling cross-modal alignment of visual and textual data. This paper presents a novel framework that combines vertical federated learning with CLIP to enable privacy-preserving medical image analysis across institutional boundaries. Our framework allows secure analysis of distributed medical data without raw data sharing, while optimizing CLIP's performance for medical applications through Context Optimization. Experimental validation on a dataset of 7023 brain MRI scans demonstrates the framework's effectiveness, achieving 93.1% accuracy in classifying four types of brain conditions (glioma, meningioma, pituitary, and no tumor) - a substantial improvement from the original pre-trained CLIP model's 26.3% accuracy. These results establish a practical solution for secure, cross-institutional medical data analysis that maintains patient privacy while enabling critical business decisions in healthcare, insurance, and financial sectors.

Keywords: Vertical federated learning · Vision-language model · Cross-institutional analysis · Prompt tuning

1 Introduction

In modern healthcare, the ability to analyze medical data across institutions holds immense potential for improving diagnostic accuracy and treatment outcomes. However, this potential is constrained by strict privacy requirements and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 320–331, 2025.
https://doi.org/10.1007/978-3-031-96997-3_20

data protection regulations such as GDPR [18] and HIPAA [19]. A particularly challenging scenario arises when third-party organizations, such as insurance companies and banks, need to assess client-related risks based on medical data distributed across various healthcare institutions.

Consider an insurance company evaluating coverage for a potential client or a bank assessing a loan application. These organizations require comprehensive medical information - including imaging data, specialist reports, and test results - from multiple healthcare providers to make informed decisions. However, healthcare institutions are bound by strict confidentiality agreements and cannot directly share raw patient data, creating a significant barrier to efficient risk assessment and decision-making.

Recent developments in artificial intelligence, particularly vision-language models like Contrastive Language-Image Pre-training (CLIP), have demonstrated exceptional capabilities in understanding relationships between visual and textual information [1]. However, deploying such models in healthcare faces two critical challenges: protecting patient privacy when analyzing data across multiple institutions, and effectively adapting general-purpose AI models to specialized medical tasks.

Our work addresses these challenges by introducing a novel framework that integrates vertical federated learning [2] with a prompt-optimized CLIP model. This integration serves two key purposes: (1) enabling secure, privacy-preserving analysis of medical data across institutional boundaries, and (2) bridging the domain gap between general-purpose vision-language models and specialized medical tasks through prompt optimization.

The proposed framework leverages the Context Optimization (CoOp) method [3] to fine-tune CLIP's text prompts specifically for medical image analysis while maintaining data privacy through vertical federated learning (VFL). This approach allows different medical institutions to participate in collaborative analysis without compromising patient privacy or regulatory compliance, while enabling third-party organizations to make informed decisions based on processed results rather than raw data.

We validate our framework through extensive experiments on brain tumor classification using a dataset of 7023 MRI scans. The results demonstrate significant improvements in classification accuracy while maintaining data privacy, suggesting a promising direction for developing secure and effective medical AI systems that can serve both healthcare and business needs.

The main **contributions** of this paper are:

- A novel framework integrating vertical federated learning with prompt-optimized CLIP for privacy-preserving medical image analysis.
- Implementation of Context Optimization (CoOp) within vertical federated learning for medical applications.
- Experimental validation demonstrating substantial improvements in brain tumor classification accuracy.

The remainder of this paper is organized as follows: Sect. 2 reviews related work in federated learning and vision-language models in the field of medi-

cal image analysis. Section 3 details the proposed framework and methodology. Section 4 presents experimental results and analysis. Finally, Sect. 5 concludes with discussions of implications and future research directions.

2 Related Works

2.1 Federated Learning

Federated learning has emerged as a promising paradigm for privacy-preserving machine learning in healthcare settings. Unlike traditional centralized learning approaches that require data aggregation, federated learning enables model training across distributed datasets while keeping sensitive data localized [[4, 20]]. Two main categories of federated learning have been established: horizontal federated learning (HFL) and vertical federated learning (VFL).

Horizontal federated learning addresses scenarios where different institutions have the same feature space but different sample sets. For example, multiple hospitals might have similar types of medical images but from different patients. Several studies have demonstrated HFL's effectiveness in medical image analysis, including chest CT classification [5], brain tumor detection [6], and retinal disease diagnosis [7].

Vertical federated learning, which is more relevant to our work, handles scenarios where different parties hold different features of the same samples. This approach is particularly suitable for healthcare scenarios where various institutions hold different types of patient data (e.g., imaging data, clinical notes, laboratory results) [8]. Recent work has shown VFL's potential in combining heterogeneous medical data sources while maintaining privacy compliance [9]. Hardy et al. [10] proposed a secure protocol for vertically partitioned data using homomorphic encryption. However, the application of VFL in the field of medical image analysis is still in its early stages [21].

2.2 Multi-modal Models in Medical Image Analysis

Recent years have witnessed growing interest in multimodal medical data mining, which aims to leverage both visual and textual information for improved diagnosis and treatment. Shetty et al. [11] proposed a multimodal deep learning framework that combines radiology images with clinical reports for more accurate pulmonary abnormality prediction. Liu et al. [12] developed an attention-based architecture that integrates pathology images with genetic data for breast cancer diagnosis. Dautov et al. [13] introduced a hierarchical learning approach that fuses medical imaging data with electronic health records (EHRs) for comprehensive patient analysis. These works demonstrated that incorporating multiple data modalities can significantly enhance medical diagnostic accuracy compared to single-modality approaches.

The emergence of CLIP has brought new opportunities to medical multimodal learning. MedCLIP [14] adapted the CLIP architecture for medical domains through specialized pre-training on medical image-text pairs, achieving

promising results in various diagnostic tasks. PMC-CLIP [15] further extended this approach by incorporating domain knowledge from medical literature, demonstrating improved performance in rare disease identification. Thawkar et al. [16] conducted a comprehensive evaluation of CLIP in medical visual question answering through PubMedCLIP, demonstrating significant but variable benefits across different medical specialties and question types, with particularly strong performance improvements in radiology and dermatology domains.

However, these models often face challenges when adapting to specific medical tasks due to the domain gap between general and medical data. To address this issue, prompt tuning methods have been proposed. The Context Optimization (CoOp) method [3] offers a particularly effective solution by learning continuous context vectors that act as optimizable prompt tokens, enabling better adaptation to medical tasks without modifying the underlying model architecture.

While these approaches have shown impressive results, they typically require centralized access to medical data, raising significant privacy concerns in real-world applications. Medical data is often distributed across different institutions (e.g., hospitals, insurance companies, research centers), each holding different aspects of patient information. Direct sharing of such sensitive data may violate privacy regulations and patient confidentiality agreements. Our work addresses this limitation by deploying CLIP within a vertical federated learning framework, enabling secure multimodal medical analysis while maintaining data privacy. This approach allows different institutions to collaborate in model training without sharing raw patient data, while leveraging CLIP's powerful vision-language capabilities through optimized prompting strategies.

3 Methodology

Our framework integrates vertical federated learning with CLIP's vision-language architecture while utilizing context optimization for medical domain adaptation. Here we detail each component and its integration into a cohesive privacy-preserving system.

Vertical Federated Learning. VFL addresses scenarios where multiple institutions hold different feature sets for the same set of patients. Unlike Horizontal Federated Learning (HFL), which assumes identical feature spaces across participants, VFL enables collaborative model training on vertically partitioned data. The foundation of our framework is built on vertical federated learning, where we establish collaboration between two distinct parties: an active party (typically an insurance company, bank, or medical center) that holds medical reports and text data, and a passive party (typically a hospital) that maintains medical imaging data. This separation reflects real-world scenarios where patient data is distributed across different healthcare institutions.

In our VFL architecture, the active party orchestrates the training process while the passive party participates by providing feature vectors from medical images. This setup ensures that raw data never leaves its source institution, with

only encrypted feature vectors and necessary gradients being exchanged between parties. The separation of data and computation aligns with privacy regulations while enabling collaborative model training.

3.1 Contrastive Language-Image Pre-training

At the core of our framework lies CLIP's dual-encoder architecture, which processes visual and textual data through separate pathways. The image encoder, residing with the passive party, processes medical images through a Vision Transformer architecture to generate fixed-dimension feature vectors. Meanwhile, the text encoder at the active party processes medical text data to generate corresponding text embeddings.

For zero-shot inference, CLIP leverages textual prompts to classify images. Given an input MRI scan x, its feature vector f is extracted by the image encoder. Simultaneously, the text encoder generates K class-specific embeddings $\{w_i\}_{i=1}^{K}$ from prompts like "a photo of a {class}", where "{class}" is replaced by medical labels (e.g., "glioma," "no tumor"). The prediction probability for class y is computed as:

$$p(y=i|x) = \frac{exp(cos(w_i, f)/\tau)}{\sum_{j=1}^{K} exp(cos(w_j, f)/\tau)} \tag{1}$$

where $cos(\cdot,\cdot)$ denotes cosine similarity, and τ is a learned temperature parameter scaling the logits.

3.2 Context Optimization

To bridge the domain gap, we integrate CoOp, a prompt-tuning method that learns a set of continuous context vectors $V = [v_1, v_2, ..., v_M]$ as prefix tokens for text prompts. For a class label 'glioma', the prompt becomes:

$$t_i = [v_1; v_2; ...;' glioma'], \tag{2}$$

where $v_i \epsilon \mathbb{R}^d$ are learnable parameters. CoOp optimizes these vectors using backpropagation, aligning CLIP's text embeddings with medical image features. Unlike full fine-tuning, this approach preserves CLIP's pre-trained knowledge while adapting to specialized tasks.

CoOp employs a cross-entropy loss to optimize the classification performance, deviating from CLIP's original contrastive loss. Given an input image x with ground-truth label y , the classification probability for class c is computed as:

$$p(y=i|x) = \frac{exp(cos(f_{text}(t_i), f_{img}(x))/\tau)}{\sum_{j=1}^{K} exp(cos(f_{text}(t_j), f_{img}(x))/\tau)} \tag{3}$$

where f_{text} and f_{img} denote the text encoder and image encoder separately. The cross-entropy loss is then defined as Eq. 4.

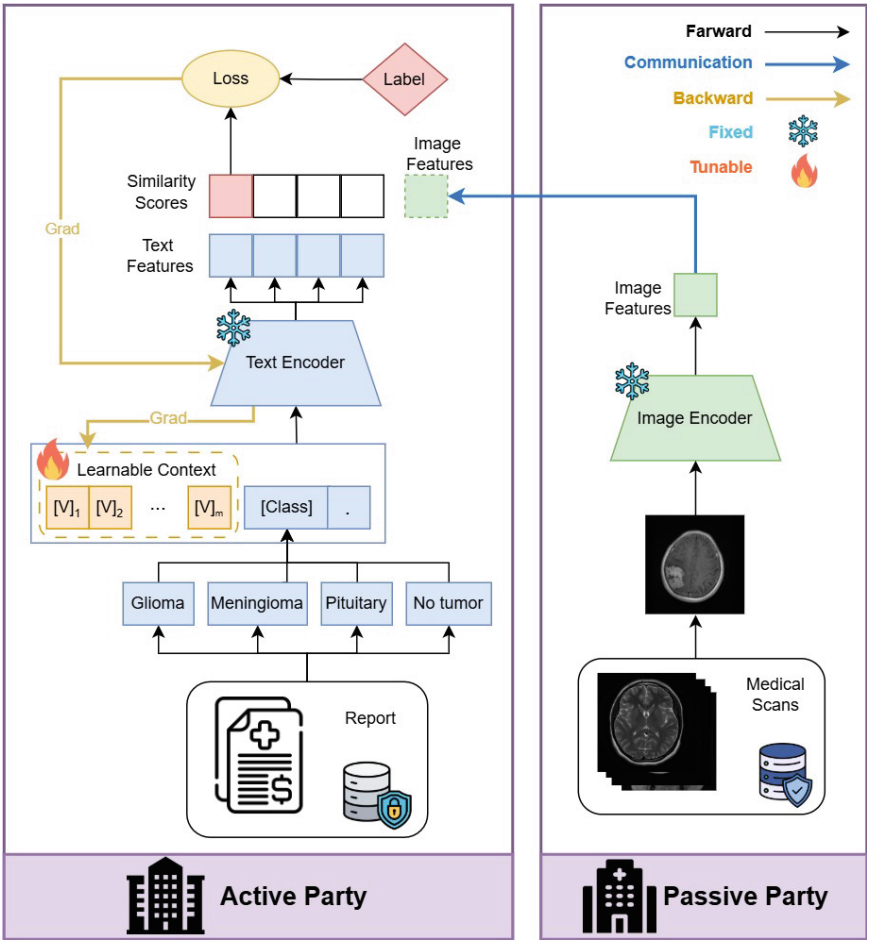

Fig. 1. Diagram of our proposed framework.

$$\mathcal{L}_{CoOp} = -\sum_{j=1}^{K} y_j \log p(t_j|x) \tag{4}$$

where y_j is the one-hot encoded ground-truth label. This loss directly optimizes the alignment between image features and task-specific textual prompts, enabling precise adaptation to medical classification tasks.

3.3 Proposed Framework

As illustrated in Fig. 1, the proposed framework integrates vertical federated learning (VFL), the CLIP model, and Context Optimization (CoOp) into a

unified pipeline designed for privacy-preserving cross-institutional medical image analysis. In this setup, medical institutions such as hospitals retain ownership of brain MRI scans, while third-party organizations (e.g., insurance companies) manage textual metadata, including clinical reports (In our work, we use diagnostic labels for simulation). This vertical partitioning ensures that raw data remains localized and is never directly shared between entities, adhering to strict privacy regulations.

The framework begins with initializing specialized models of each participant: hospitals deploy the image encoder component of CLIP, typically a Vision Transformer (ViT), to process medical images into feature embeddings, while third parties initialize the text encoder alongside learnable CoOp prompt vectors. During federated training, hospitals compute image embeddings from MRI scans locally, and third parties generate text embeddings using dynamically optimized prompts that incorporate medical terminology (e.g., "glioma" or "pituitary tumor"). These embeddings are securely aggregated through privacy-preserving protocols, such as homomorphic encryption, to compute similarity scores in a shared embedding space. The contrastive learning objective aligns image and text representations across institutions, enabling collaborative model updates without exposing sensitive raw data.

For inference, new MRI scans are classified by comparing their image embeddings against the optimized text prompts derived from CoOp, which have been tuned to capture domain-specific medical features. This combination of VFL's distributed architecture, CLIP's multimodal alignment capabilities, and CoOp's adaptive prompting establishes a robust solution for cross-institutional medical analysis that balances diagnostic accuracy with stringent data confidentiality.

4 Experiments and Results

4.1 Experimental Setup

Dataset. We conducted our experiments using a comprehensive brain tumor MRI dataset [17] containing 7023 images across four classes: glioma, meningioma, pituitary tumors, and no tumor (healthy control). The dataset is publicly available on Kaggle and is balanced across the four classes, with each class having approximately equal numbers of images. Figure 2 shows sample images from each class for reference. To simulate a realistic federated learning scenario, we distributed the data between two parties: the passive party (hospital) holding MRI images and the active party (medical center) holding diagnostic labels.

Configuration. We implemented our framework using PyTorch [22] and used the ViT-L/14@336px architecture as our base model. The learnable context consisted of 10 trainable vectors, each with 512 dimensions. All experiments were conducted on a server equipped with NVIDIA A800 GPUs, with secure communication protocols implemented between parties. For optimization, we used the Adam optimizer with a learning rate of 0.001 for the context vectors, while keeping the CLIP model parameters frozen. Training was conducted for 100

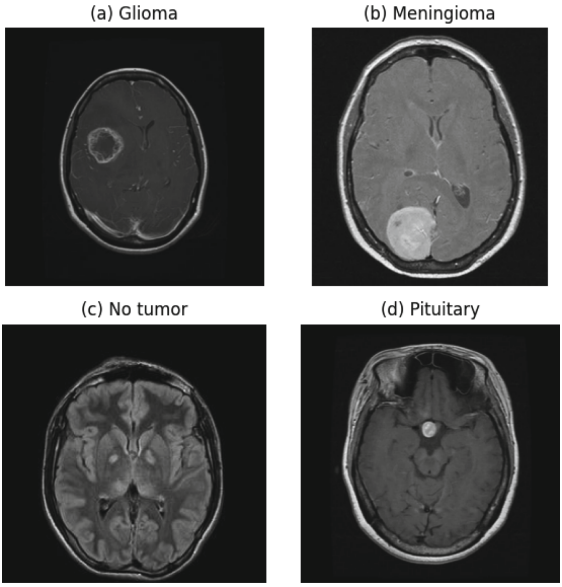

Fig. 2. Sample MRI scans.

epochs with a batch size of 32. We employed early stopping based on validation performance to prevent overfitting.

Baseline Models. We compared the performance of our framework against several baselines, including:

- Original CLIP models pretrained on general data: clip_ViT_L_14_336, clip_ViT_B_16, clip_ViT_B_32;
- PubMed-CLIP, a domain-specific CLIP model pretrained on medical data.

Evaluation Metrics. To comprehensively evaluate our framework's performance, we employed several standard metrics for multi-class classification:

- Accuracy: The ratio of correctly predicted observations to the total observations, providing an overall measure of model performance.
- Precision: For each class, the ratio of true positive predictions to the total positive predictions (true positives + false positives), indicating the model's ability to avoid false positive classifications.
- Recall: For each class, the ratio of true positive predictions to the total actual positives (true positives + false negatives), measuring the model's ability to find all positive cases.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.
- Macro F1-score: The unweighted average of F1-scores for each class, giving equal importance to each class regardless of its support.

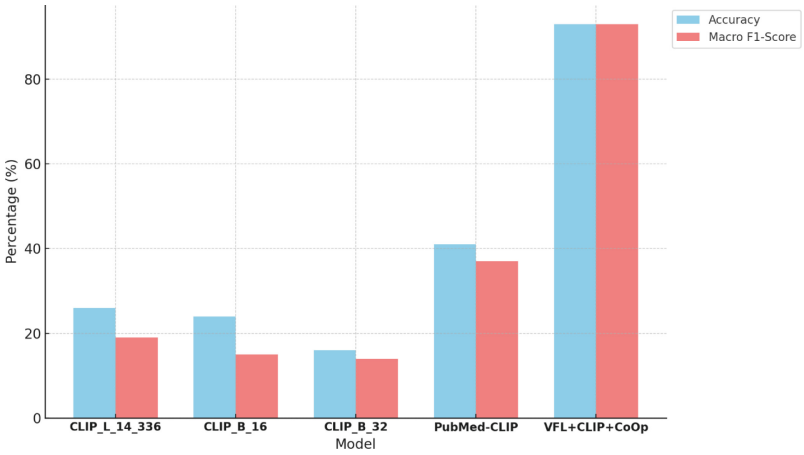

Fig. 3. Models Performance Comparison.

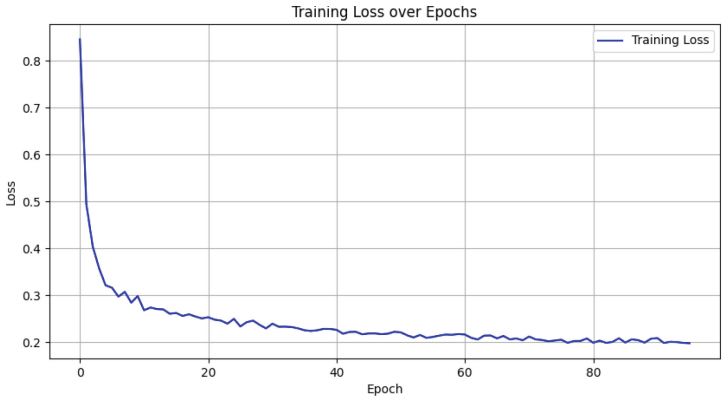

Fig. 4. Loss of our proposed framework.

These metrics were chosen because accuracy alone can be misleading in medical applications where the cost of different types of errors may vary significantly. For instance, misclassifying a tumor type could lead to inappropriate treatment decisions. The combination of precision, recall, and F1-score provides a more nuanced understanding of the model's performance across different tumor classes.

4.2 Results

Our experimental results demonstrated significant performance improvements over the baselines. Figure 3 illustrates the comparative accuracy and macro

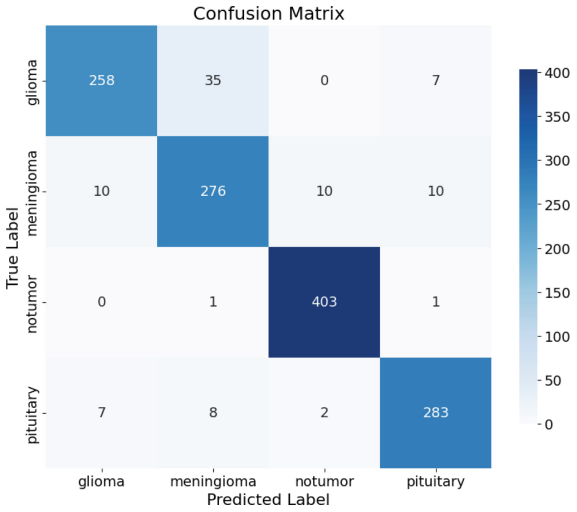

Fig. 5. Confusion matrix of our proposed framework.

F1-scores of different models. The original pretrained CLIP models exhibit relatively low performance, with accuracy ranging from 16% to 26% and macro F1-scores between 14% and 19%. In contrast, PubMed-CLIP, which is pretrained on medical data, shows moderate improvement with an accuracy of 41% and a macro F1-score of 37%. However, our proposed framework, which integrates vertical federated learning and CLIP model with CoOp optimization, dramatically outperforms these models, achieving 93.1% in both accuracy and macro F1-score. The sharp increase in both accuracy and F1-scores validates the effectiveness of our privacy-preserving framework in real-world medical scenarios, where data is distributed across institutions.

The loss values during training of our framework are illustrated in Fig. 4. It shows a clear downward trend in the loss as the epochs progress, stabilizing after approximately 80 epochs. This behavior indicates the model's ability to learn effectively without overfitting, as demonstrated by the early stopping mechanism triggered based on validation performance.

Additionally, we analyzed the classification performance for each individual class: glioma, meningioma, pituitary tumors, and no tumor. Figure 5 illustrates the confusion matrix for the classification results of our proposed framework. The matrix demonstrates a balanced performance across the different tumor types, with minimal misclassifications. Table 1 provides the detailed precision, recall, and F1-scores for each class, showcasing the robust performance of the proposed method across all tumor types. Notably, the "no tumor" class achieves exceptionally high precision (97.1%) and recall (99.5%), reflecting the model's ability to accurately distinguish healthy controls from tumor cases.

Table 1. Classification Metrics of Our Proposed Framework.

Class Name	Precision	Recall	F1-score
Glioma	93.8%	86.0%	89.7%
Meningioma	86.3%	90.2%	88.2%
No Tumor	97.1%	99.5%	98.3%
Pituitary	94.0%	94.3%	94.2%

5 Conclusion

In this paper, we proposed a novel privacy-preserving framework for cross-institutional medical image analysis by combining vertical federated learning (VFL) with the Contrastive Language-Image Pretraining (CLIP) model, enhanced through Context Optimization (CoOp). This framework facilitates secure collaboration across medical institutions while safeguarding patient data. Our results show significant improvements in brain tumor classification, achieving 93.1% in both accuracy and macro F1-score, outperforming original CLIP variants and PubMed-CLIP.

By using VFL, the framework addresses domain adaptation challenges in medical image analysis and enables collaboration without raw data sharing. Additionally, aligning medical images with optimized text prompts enhances diagnostic accuracy by bridging general-purpose vision-language models with specialized medical tasks.

Limitations and Future Work. Future research could explore incorporating richer textual data, such as clinical notes, to improve model interpretability. Expanding the framework to other medical domains with diverse datasets and integrating heterogeneous data modalities (e.g., EHRs, genomic data) would further enhance diagnostic capabilities.

In summary, the proposed framework offers a scalable solution for privacy-preserving medical image analysis, balancing data privacy with diagnostic accuracy to enable secure collaboration between healthcare institutions.

References

1. Radford, A., et al.: Learning transferable visual models from natural language supervision. Int. Conf. Mach. Learn. 8748–8763 (2021)
2. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**, 1–19 (2019)
3. Zhou, K., Yang, J., Loy, C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**, 2337–2348 (2022)
4. Li, T., Sahu, A., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. **37**, 50–60 (2020)

5. Dou, Q., et al.: Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ Digital Med. **4**, 60 (2021)
6. Sheller, M., Reina, G., Edwards, B., Martin, J., Bakas, S.: Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke And Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held In Conjunction With MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4, pp. 92–104 (2019)
7. Gulati, S., Guleria, K., Goyal, N., AlZubi, A., Castilla, Á.: A Privacy-Preserving Collaborative Federated Learning Framework for Detecting Retinal Diseases. IEEE Access (2024)
8. Yan, Y., et al.: Cross-modal vertical federated learning for MRI reconstruction. IEEE J. Biomed. Health Inf. **28**, 6384–6394 (2024)
9. Khan, A., Thij, M., Wilbik, A.: Vertical federated learning: a structured literature review. Knowl. Inf. Syst. 1–39 (2025)
10. Hardy, S., et al.: Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. ArXiv Preprint ArXiv:1711.10677 (2017)
11. Shetty, S., Ananthanarayana, S., Mahale, A.: Multimodal medical tensor fusion network-based DL framework for abnormality prediction from the radiology CXRs and clinical text reports. Multimedia Tools Appl. **82**, 44431–44478 (2023)
12. Liu, H., Shi, Y., Li, A., Wang, M.: Multi-modal fusion network with intra-and inter-modality attention for prognosis prediction in breast cancer. Comput. Biol. Med. **168**, 107796 (2024)
13. Dautov, R., Distefano, S., Buyya, R.: Hierarchical data fusion for smart healthcare. J. Big Data **6**, 1–23 (2019)
14. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2022, pp. 3876 (2022)
15. Lin, W., et al.: PMC-CLIP: contrastive language-image pre-training using biomedical documents. In: International Conference On Medical Image Computing And Computer-Assisted Intervention, pp. 525–536 (2023)
16. Eslami, S., Meinel, C., De Melo, G.: PubMedCLIP: how much does CLIP benefit visual question answering in the medical domain? Find. Assoc. Comput. Linguist. EACL **2023**, 1181–1193 (2023)
17. Brain Tumor MRI Dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. Accessed 25 Feb 2025
18. Regulation, P. Regulation (EU) 2016/679 of the European Parliament and of the Council. Regulation (eu). **679**, 2016 (2016)
19. Act, A.: Health insurance portability and accountability act of 1996. Public Law. **104**, 191 (1996)
20. Chen, Y., Qin, X., Wang, J., Yu, C., Gao, W.: FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell. Syst. **35**, 83–93 (2020)
21. Sohan, M., Basalamah, A.: A systematic review on federated learning in medical image analysis. IEEE Access **11**, 28628–28644 (2023)
22. Paszke, A.: PyTorch: An imperative style, high-performance deep learning library. ArXiv Preprint ArXiv:1912.01703. (2019)