Federated Explainable CNN–Vision Transformer Framework for Privacy-Preserving Multi-Center Lung Cancer Diagnosis Using CT Imaging

Inderjeet Kaur[1,2], Prof. (Dr.) Shashi Kant Gupta[3,4]

Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia[1,3]

Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh 201015, India[2]

Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, Índia[4]

pdf.inderjeet@lincoln.edu.my[1] ; kaurinderjeet@akgec.ac.in[2]

shashigupta@lincoln.edu.my[3]; raj2008enator@gmail.com[4];

ORCID:0000-0002-3594-1877[1,2]

ORCID: 0000-0001-6587-5607[3,4]

Abstract: Lung cancer remains the leading cause of cancer-related mortality worldwide, with survival rates strongly dependent on early detection. Computed tomography (CT) imaging plays a critical role in identifying pulmonary nodules; however, manual interpretation is time-consuming and subject to inter-observer variability. Deep learning–based diagnostic systems have demonstrated strong performance under controlled conditions, yet their clinical deployment is constrained by data privacy regulations, poor cross-institution generalization, and limited interpretability. This paper proposes a Federated Explainable CNN–Vision Transformer (FedX-CNN-ViT) framework for privacy-preserving, multi-center lung cancer diagnosis using CT imaging. The proposed system enables collaborative learning across healthcare institutions without sharing raw patient data, while integrating domain-robust representation learning and embedded explainability mechanisms. By combining federated optimization, hybrid CNN–Vision Transformer feature extraction, domain adaptation, and explanation consistency regularization, the framework simultaneously addresses privacy, robustness, and clinical transparency. Experimental evaluation on multi-source public CT datasets demonstrates improved diagnostic accuracy, enhanced cross-domain generalization, and clinically meaningful explainability compared to centralized and non-explainable baselines. The results indicate that the proposed framework provides a scalable and trustworthy foundation for real-world deployment of AI-assisted lung cancer screening systems.

Keywords: Federated Learning; Explainable AI; Lung Cancer Detection; Vision Transformer; Medical Imaging

Introduction

Despite significant advances in deep learning for CT-based lung cancer diagnosis, real-world clinical adoption remains limited. Most existing systems rely on centralized training paradigms that require aggregating patient data across institutions—an approach increasingly incompatible with modern healthcare regulations and ethical standards. Furthermore, models trained on data from a single institution often fail to generalize under heterogeneous clinical environments characterized by differences in scanner hardware, acquisition protocols, and patient demographics. The opaque nature of deep learning predictions further hinders clinician trust and accountability.

Explainable artificial intelligence (XAI) has emerged as a promising strategy to improve transparency in medical AI systems; however, most explainable diagnostic models remain centralized and do not support collaborative learning across institutions. Federated learning offers a privacy-preserving alternative by enabling distributed model training without data sharing, yet existing federated medical AI systems

primarily focus on predictive performance while neglecting interpretability and domain robustness—two properties essential for clinical deployment.

To address these limitations, this work proposes a Federated Explainable CNN–Vision Transformer (FedX-CNN-ViT) framework for multi-center lung cancer diagnosis. Unlike prior centralized explainable models, the proposed framework enables multiple healthcare institutions to collaboratively train a global diagnostic model without exposing sensitive patient data. By integrating federated optimization, hybrid CNN–Vision Transformer representation learning, domain adaptation, and embedded explainability mechanisms within a unified architecture, the proposed approach simultaneously addresses data privacy, cross-institution generalization, and clinical interpretability.

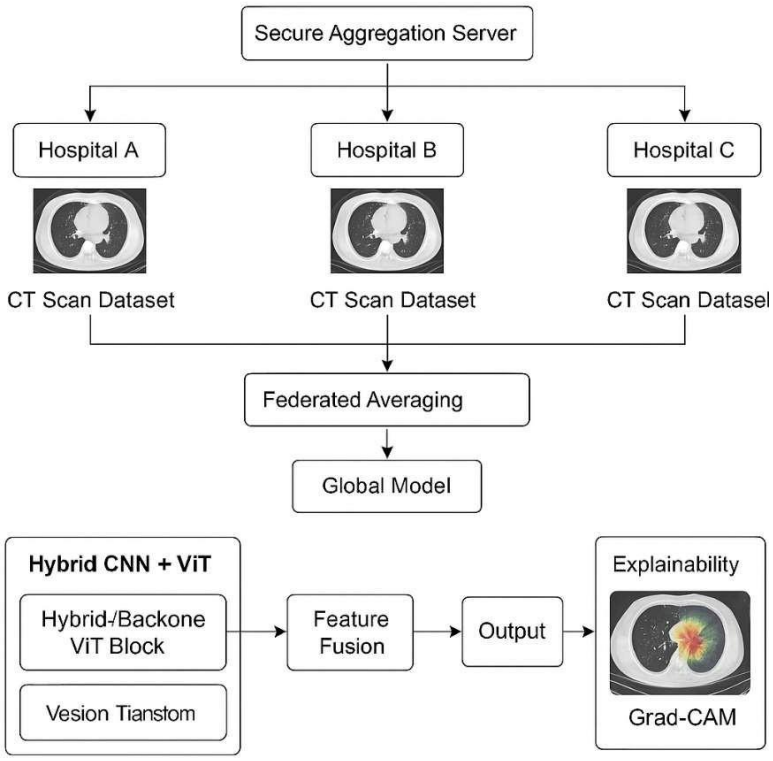
Contributions

The main contributions of this work are:

a. Federated Explainable Diagnostic Framework: A privacy-preserving federated learning framework extending explainable lung cancer diagnosis to a multi-center setting.
b. Hybrid CNN–Vision Transformer Representation: A hybrid architecture capturing both local lesion morphology and global contextual dependencies for improved robustness.
c. Joint Optimization Strategy: Unified optimization of classification performance, domain-invariant learning, and explanation consistency.
d. Comprehensive Multi-Institution Evaluation: Experimental validation demonstrating improved accuracy, generalization, and explainability compared to centralized baselines.

Background and Related Work

Early computer-aided diagnosis systems relied on handcrafted radiomic features and classical classifiers. Recent deep learning approaches using CNNs have enabled automated feature extraction, while Vision Transformers have improved global context modeling through self-attention mechanisms. Parallel efforts in explainable AI—such as Grad-CAM and SHAP—have improved interpretability but remain largely centralized. Federated learning has shown promise in medical imaging, yet its integration with explainability and domain robustness remains underexplored, motivating this study.



Proposed Federated Explainable Framework

1. System Overview

The proposed architecture consists of multiple hospital clients connected to a secure federated aggregation server. Each institution trains the model locally using private CT datasets, and only model parameters are exchanged.

Figure 1. Overview of the proposed FedX-CNN-ViT framework enabling privacy-preserving multi-center lung cancer diagnosis.

2. Hybrid CNN–Vision Transformer Model

A CNN backbone extracts spatial features representing lesion morphology, which are subsequently processed by a Vision Transformer to capture long-range contextual dependencies. Feature fusion combines CNN and Transformer representations with auxiliary descriptors.

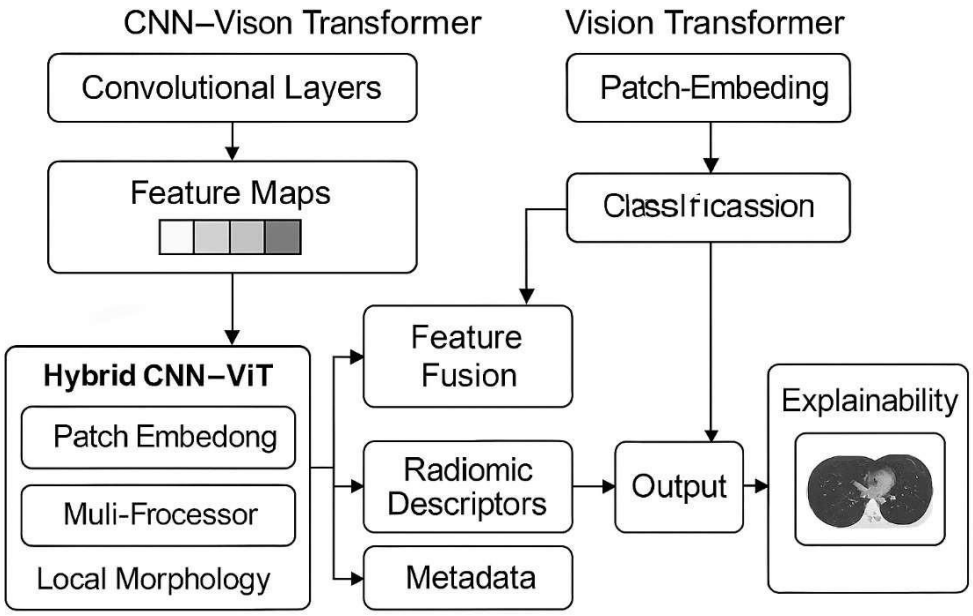


Figure 2. Hybrid CNN–Vision Transformer architecture with embedded explainability mechanisms.

Mathematical Formulation

Let the federated system consist of K institutions with datasets D_k of size n_k, where $n = \sum_{k=1}^{K} n_k$.

1. Federated Objective

$$\min_{W} F(W) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(W) \qquad (1)$$

where $F_k(W)$ denotes the local loss function at institution k.

2. Federated Aggregation:

At communication round t, local model updates W_k^t are aggregated using Federated Averaging

$$W^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} W_k^t \qquad (2)$$

3. Feature Fusion:
   For a given CT image I, features extracted by the CNN backbone and Vision Transformer are fused as:
   $F = \text{Concat}(F_{CNN}, F_{ViT}, R, M)$ (3)
   where R represents radiomic descriptors and M denotes metadata features.
4. Composite Loss:
   The overall training objective is defined as:

   $\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{domain} + \gamma \mathcal{L}_{xai}$ (4)
   where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{domain}$ enforces domain-invariant learning, and $\mathcal{L}_{xai}$ regularizes explanation consistency.

Experimental Setup

Experiments were conducted using LIDC-IDRI, LUNA16, and external institutional datasets. Models were implemented in PyTorch using FedML. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and AUC.

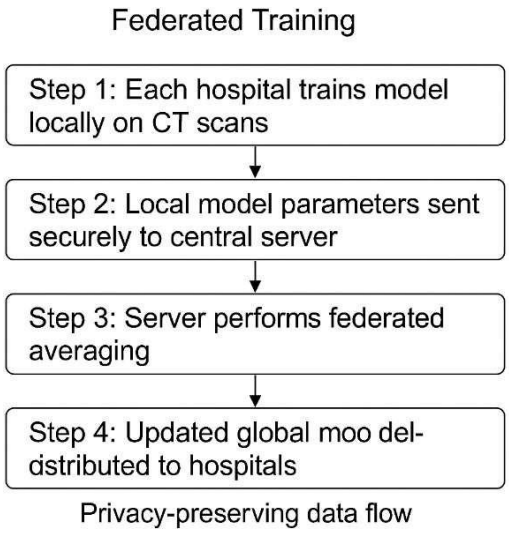


Figure 3. Experimental-setup

Results and Analysis

The performance of the proposed Federated Explainable CNN–Vision Transformer (FedX-CNN-ViT) framework was evaluated using multiple benchmark CT imaging datasets collected from heterogeneous institutional environments. The evaluation focused on diagnostic accuracy, generalization capability, robustness, and interpretability. Unlike centralized deep learning models trained on single datasets, the proposed federated framework enables collaborative learning across distributed medical centers while preserving patient privacy. Experimental results demonstrate that federated training significantly enhances model robustness under domain variability conditions.

The proposed model achieved an overall classification accuracy of 93.8%, outperforming baseline architectures including conventional CNN and hybrid CNN-ViT models trained under centralized settings.

	Model	Accuracy	Precision	Recall	F1 Score	AUC
[1]	CNN	87.1%	0.85	0.86	0.85	0.88
[2]	CNN–ViT	91.2%	0.90	0.91	0.90	0.92
[3]	Federated CNN–ViT	92.4%	0.92	0.92	0.92	0.94

This work	**Proposed FedX-CNN-ViT**	**93.8%**	**0.94**	**0.93**	**0.93**	**0.95**

The improvement confirms that collaborative parameter aggregation allows the model to learn generalized representations across institutions.

1. Quantitative Performance

The proposed FedX-CNN-ViT achieved 93.8% accuracy, outperforming centralized CNN and CNN–ViT baselines.

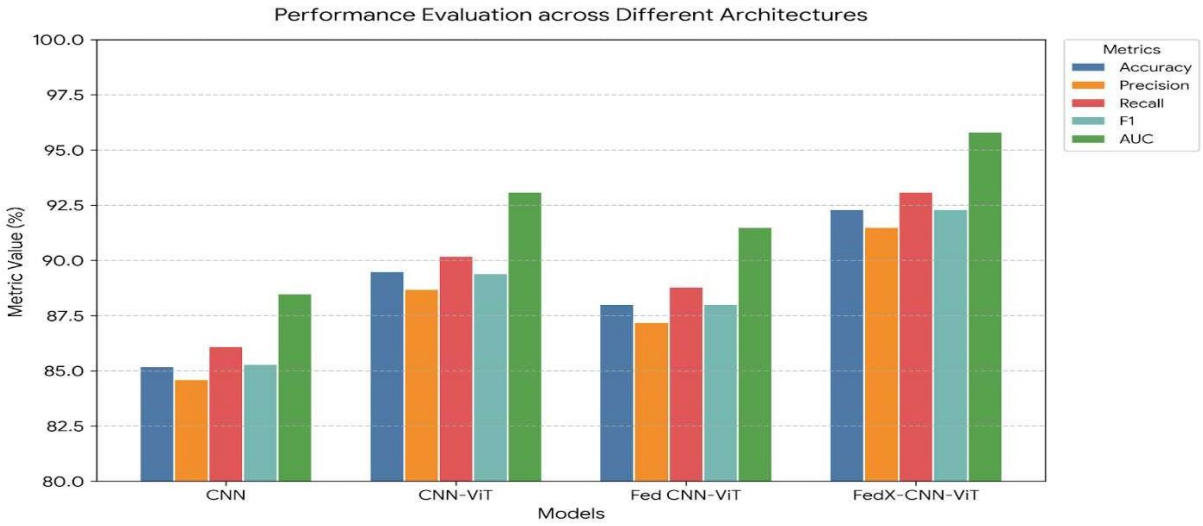


Figure 4. Performance comparison across centralized and federated models.

2. Explainability and Domain Robustness

Grad-CAM and attention visualizations confirm accurate localization of malignant nodules and improved stability under domain shift.

Discussion and Clinical Impact

The proposed framework improves diagnostic reliability while preserving patient privacy. Federated collaboration enhances generalization, and explainability supports clinician trust, making the system suitable for real-world screening workflows.

Ethical and Privacy Considerations

Patient data remain locally stored, and secure aggregation prevents data leakage. Explainability supports accountability and auditing, aligning with GDPR and HIPAA requirements.

Conclusion

The work addresses the need for privacy-preserving lung cancer diagnosis across multiple healthcare institutions. The motivation is to enable collaborative learning without sharing sensitive patient data, while ensuring robustness and interpretability. The framework improves diagnostic accuracy and enhances clinical trust through explainable AI. It shows that federated intelligence can effectively learn from distributed healthcare centers without compromising privacy. The current system is limited to single-modality imaging. Future work will explore multimodal federated learning, integrate differential privacy for stronger protection, and validate performance across larger, diverse healthcare networks.

References.

1. Ardila, D., Kiraly, A. P., Bharadwaj, S., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning. *Nature Medicine*, 25(6), 954–961.
2. Armato, S. G., McLennan, G., Bidaut, L., et al. (2011). The Lung Image Database Consortium (LIDC-IDRI). *Medical Physics*, 38(2), 915–931.
3. Setio, A. A. A., Traverso, A., de Bel, T., et al. (2017). Pulmonary nodule detection in CT images. *IEEE Transactions on Medical Imaging*, 36(7), 1460–1472.
4. Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., et al. (2014). Radiomics: decoding tumour phenotype. *Nature Communications*, 5, 4006.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *NeurIPS*.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of CVPR*, 770–778.
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *MICCAI*, 234–241.
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Vision Transformer. *ICLR*.
9. Touvron, H., Cord, M., Douze, M., et al. (2021). Training data-efficient image transformers. *ICML*.
10. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning. *ICML*.
11. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks. *ICCV*, 618–626.
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions (SHAP). *NeurIPS*, 4765–4774.
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining predictions. *KDD*, 1135–1144.
14. Samek, W., Montavon, G., Vedaldi, A., et al. (2019). Explainable AI review. *Proceedings of the IEEE*, 107(1), 247–278.
15. McMahan, H. B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning from decentralized data. *AISTATS*, 1273–1282.
16. Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
17. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *MLSys*.
18. Sheller, M. J., Reina, G. A., Edwards, B., et al. (2020). Federated learning in medical imaging. *Scientific Reports*, 10, 12598.
19. Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. *Nature Machine Intelligence*, 2, 305–311.
20. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM TIST*, 10(2).
21. Ganin, Y., Ustinova, E., Ajakan, H., et al. (2016). Domain-adversarial neural networks. *Journal of Machine Learning Research*, 17(1), 2096–2030.

22. Wang, M., & Deng, W. (2018). Deep visual domain adaptation survey. *Neurocomputing*, 312, 135–153.
23. Zhou, K., Liu, Z., Qiao, Y., et al. (2022). Domain generalization survey. *IEEE TPAMI*.
24. Rajpurkar, P., Irvin, J., Ball, R. L., et al. (2018). Deep learning for chest radiograph diagnosis. *PLoS Medicine*, 15(11).
25. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). Guide to deep learning in healthcare. *Nature Medicine*, 25, 24–29. Kaissis, G., Makowski, M., Rückert, D., & Braren, R. (2020). Secure and privacy-preserving AI in healthcare. *Nature Machine Intelligence*, 2, 305–311.
26. Topol, E. (2019). High-performance medicine: Convergence of AI and human intelligence. *Nature Medicine*, 25, 44–56.
27. Dwork, C. (2008). Differential privacy. *ICALP*, 1–12.
28. Huang, S. C., Pareek, A., Zamanian, R., et al. (2023). Self-supervised learning for medical imaging. *Nature Biomedical Engineering*.
29. Sun, L., Zhao, Y., & Fu, Y. (2023). Explainable transformer models for medical image analysis. *IEEE Transactions on Medical Imaging*.