Contents available at ITME Online Repository

INNOVATIVE TRENDS IN MULTIDISCIPLINARY ENGINEERING

International Peer-Reviewed Journal of the School of Engineering & Technology

K.R. Mangalam University, Gurugram

Journal Homepage: **https://itme.krmangalam.edu.in/**



FedXAI-Health: Privacy-Protected Explainable Federated Deep Learning Model in Pre-emption of Early Diseases with Multimodal Healthcare Data

Kaljot Sharma[a], **Neetu Rani**[a,*]

[a] *Department of CSE, Chandigarh University, Mohali, India*

DOI: 10.68343/ITME.2026.0102.5

ARTICLE INFO

Article history:
Received 23 March 2026
Revised 26 April 2026
Accepted 15 May 2026
Available online DD Month 2026

Keywords:
Federated Learning
Explainable AI
Multimodal Learning
Healthcare Analytics
Deep Learning
Privacy-Preserving AI

ABSTRACT

Machine learning (ML) has significantly accelerated healthcare analytics, but its widespread clinical implementation is hindered by issues of data privacy, inter-institutional heterogeneity, and a lack of interpretability. To overcome these challenges, this paper proposes the implementation of FedXAI-Health, a novel framework that combines federated learning (FL), multimodal deep learning, and explainable artificial intelligence (XAI) to predict diseases while preserving privacy. The proposed methodology enables multiple healthcare facilities to collaboratively train a global model without transferring raw patient information. A hybrid model between convolutional neural networks (CNNs) of medical imaging and transformer-based models of electronic health records (EHRs) is used to learn complementary attributes of heterogeneous datasets. The Federated Averaging (FedAvg) algorithm is used to combine model updates to provide efficient and secure distributed training. SHAP-based explainability is also added to increase the level of transparency by offering feature-level explanations of model predictions. It is tested with benchmark data, such as MIMIC-III (EHR data) and chest X-ray images, and in a simulated federated environment. It is proved that the proposed model has better performance than the baseline techniques, with an accuracy of 94.2 per cent and an AUC of 0.96, and keeps the data private. The explainability module also delivers clinically relevant information in line with the domain knowledge. The suggested FedXAI-Health architecture provides a scalable, interpretable and privacy-sensitive solution to the next generation of smart medical systems. Further development will be centred on the clinical validation in practice and integration with real-time systems of patient monitoring.

1. Introduction

The high-rate development of machine learning (ML) and deep learning (DL) has greatly changed the sphere of healthcare, as now it is possible to detect illnesses early on, diagnose the disease predictively, and make clinical decisions based on the information. Medical image analysis, electronic health record (EHR) mining, and personalized treatment recommendation systems are examples of applications where AI is potentially useful to enhance patient outcomes and healthcare efficiency. Regardless of these developments, the realworld application of ML models to practical healthcare scenarios is a subject of limited use because of some severe issues.

Data privacy and security can be listed as one of the main challenges. Healthcare data are very sensitive, and very stringent rules normally do not allow institutions to provide patient-level data to different organizations.

* Corresponding author. E-mail: **neetuvashist@gmail.com**

© 2026 The Authors. Published by K.R. Mangalam University, Gurugram.
Selection and double-blind peer-review under the responsibility of School of Engineering & Technology.
This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

Conventional centralized methods of learning demand the collection of information in one source, which enhances privacy violations and restricts mass cooperation. In a bid to resolve this challenge, federated learning (FL) has come into the picture as one of the potential paradigms, which can support collaborative model training without sharing raw data.

The heterogeneity of healthcare data is another very important issue. Clinical information is always multimodal, including structured (e.g., EHRs and laboratory results) as well as unstructured (e.g., medical images) data. Current methods can only process data based on single modalities, and therefore the model is unable to fully address the complexity of the patient conditions. Multimodal data integration would significantly enhance the predictive performance, yet this would add more complexity to model design and training.

Moreover, deep learning models are not interpretable, which is another significant impediment to clinical usage. Most of the state-of-the-art models are black boxes, meaning that they give predictions without giving explicit explanations. In the healthcare field, where choices are vital, clinicians need to have transparent and explainable systems to trust and legitimise model results.

Despite the existence of research into federated learning in healthcare and multimodal deep learning, as well as explainable AI, there is very limited literature that combines these concepts to achieve the same outcome: preserving privacy, fusing multimodal data, and maintaining interpretability. Current models of federated learning tend to deal with the data of a single modality and do not provide explainability, whereas multimodal models are usually trained in centralised environments, restricting their usage in privacy-sensitive environments.

To solve these shortcomings, the given research paper intends to create a privacy-conscious, multimodal, and understandable machine learning platform in healthcare prediction that will be capable of functioning in distributed clinical settings.

The current paper suggests FedXAI-Health, a new model that combines:

- **Federated learning (FL):** An example of decentralized and privacy-preserving training.
- **Multimodal Deep Learning:** to integrate EHR and medical imaging data.
- **Explainable AI (XAI):** in order to give interpretable predictions to aid clinical decision support.

The conclusions of this work can be summarized as follows:

1. **Single Framework:** We suggest an innovative design combining federated learning, multimodal deep learning, and explainable AI in medical prediction.
2. **Multimodal Feature Learning:** A hybrid system consisting of CNNs to process the imaging data and transformer-based architectures to process the EHR data is created to extract complementary features.
3. **Privacy-Preserving Training:** The framework uses the Federated Averaging (FedAvg) algorithm to facilitate the learning process of distributed healthcare institutions in a collaborative manner without exchanging raw data.
4. **Explainability Integration:** SHAP-based explanations are added to enable interpretability of features on the feature level, which increases clinical trust and usability.
5. **Overall assessment:** A thorough examination of the suggested methodology is done on benchmark data sets in a simulated federated environment and against baseline models that it outperforms.

The suggested FedXAI-Health framework uses a practical process to fill the gap between privacy preservation, multimodal learning, and interpretability with the implementation of trustworthy AI systems into real healthcare settings.

2. Literature Review

Machine learning and deep learning have found extensive application in healthcare in diagnosis, prognosis, risk prediction, and clinical decision support. Initial experiments have shown that deep models have the capacity to discover clinically significant patterns in large-scale electronic health records (EHRs). To

illustrate, Rajkomar et al. created scalable deep learning models with EHR data to predict in-hospital mortality, readmission, and long length-of-stay by demonstrating that the data-driven approach is superior in predicting conventional clinical scoring schemes [1]. Likewise, Miotto et al. proposed Deep Patient, an EHR-based representation learning, which is an unsupervised deep architecture and emphasizes the importance of deep learning to learn latent features about patients that can be used to predict diseases [2].

Simultaneously, medical imaging has turned out to be among the most effective places of deep-learning application in medicine. Using convolutional neural networks, Esteva et al. reported a dermatologist-level performance in skin cancer classification, which indicates that image-based deep learning systems have clinical potential [3]. Nonetheless, since healthcare data is usually heterogeneous, predictive performance can be constrained by a single modality. Multimodal learning has thus emerged as an effective approach. Surveys such as the one conducted by Huang et al. highlight that imaging, EHR, sensor, and genomic data can result in stronger and more informative prediction models when used together [4].

These advances notwithstanding, privacy, governance, and institutional silos of data limit the implementation of centralized machine learning in the healthcare sector. Federated learning (FL) has been highly regarded for overcoming these challenges. Rieke et al. addressed how FL can be used in transformative manners in digital health, specifically in making it possible to conduct distributed learning without transporting sensitive patient information [5]. Kaissis et al. also contended that FL, together with secure computation and privacy-preserving techniques, can be the foundation of trustworthy AI systems in medical imaging and analytics in healthcare fields [6]. Practical proof of FL performance was reported by Sheller et al., who used federated learning for brain tumor segmentation across centers and indicated that collaborative training could achieve performance comparable to the centralized one and privacy could be preserved [7].

The second important issue is that not all healthcare artificial intelligence models are interpretable. Clinical decisions need to be justified, which is why explainable artificial intelligence (XAI) has become a necessity. Tjoa and Guan conducted a review of the role of XAI in healthcare and found that black-box models, although being highly accurate, are often met with resistance among clinicians when explanations are not available [8]. Similarly, Holzinger et al. also pointed out that explainability is essential to trust, accountability, and safe AI implementation in medicine, particularly in high-stakes environments [9].

The summary of literature is shown in Table 1. Even though the development of EHR-based learning, medical imaging, federated learning, and explainable AI has been investigated separately so far, there is a lack of literature that combines all these features within one framework. It is common that existing healthcare FL research tends to concentrate on a single modality, whereas most multimodal healthcare models are formulated in centralized locations. Additionally, interpretability is often viewed as an optional complement (as opposed to being a fundamental part of the predictive pipeline). This shows an obvious gap: it requires a single framework that would equally facilitate privacy-preserving distributed learning, multimodal feature integration, and explainable features that are meaningful to the clinician. The proposed FedXAI-Health framework fills this gap by integrating federated learning and multimodal deep learning with SHAP-based interpretability into one healthcare prediction pipeline.

Table 1. Summary of Key Literature

Ref.	Study	Focus Area	Key Contribution	Limitation
[1]	Rajkomar et al.	EHR-based DL	Large-scale clinical prediction from EHRs	Centralized setting
[2]	Miotto et al.	EHR representation learning	Deep latent patient representation	Limited interpretability
[3]	Esteva et al.	Medical imaging	Dermatologist-level image classification	Single modality
[4]	Huang et al.	Multimodal healthcare AI	Review of multimodal data fusion in healthcare	Mostly survey-based
[5]	Rieke et al.	Federated learning	FL as a privacy-preserving paradigm in digital health	Conceptual, broad scope
[6]	Kaissis et al.	Privacy-preserving ML	Trustworthy AI through FL and secure methods	More imaging-oriented
[7]	Sheller et al.	FL in medical imaging	Cross-institutional tumor segmentation with FL	Task-specific

Ref.	Study	Focus Area	Key Contribution	Limitation
[8]	Tjoa and Guan	Explainable AI	Review of XAI methods in healthcare	Limited multimodal focus
[9]	Holzinger et al.	Explainability in medicine	Importance of causability and human-centered AI	Mostly conceptual
[10]	Shickel et al.	Deep learning on EHRs	Comprehensive review of DL methods for EHR	Pre-FL emphasis

3. Methodology

This paper suggests FedXAI-Health, a combined solution comprising federated learning (FL), multimodal deep learning, and explainable AI (XAI) to privacy-preserving healthcare prediction. The system has distributed client-server architecture, according to which several healthcare organizations jointly train a global model without raw patient data.

Multimodal inputs that include Electronic Health Records (EHR) and medical imaging data are processed by each client. A Convolutional Neural Network (CNN) is used to extract spatial features of imaging data and a Transformer-based model is used to extract temporal and relational patterns in structured EHR data. A feature fusion layer is used to combine the extracted representations, which allows the model to learn complementary multimodal features.

A federated learning mechanism is used in model training. Let K be the number of participating clients. All clients update the parameters of local models with their own dataset, and the central server consolidates these updates using the Federated Averaging (FedAvg) algorithm:

$$W(t) = \sum_{i=1}^{K} \frac{n_i}{N} W_i(t) \tag{1}$$

where,

- $W_i(t)$: local model weights
- n_i: number of samples at client i
- N: total samples

To enhance interpretability, the framework integrates SHAP (SHapley Additive Explanations), which quantifies feature contributions to model predictions:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{2}$$

where φ_i represents the contribution of feature i.

Due to the lack of patient-level aligned multimodal datasets, a simulated multimodal setup is adopted to evaluate the integration capability of the proposed model. This provides a controlled environment for assessing multimodal fusion performance, while future work will focus on clinically aligned datasets. Figure 1 illustrates the complete architecture of the proposed framework.

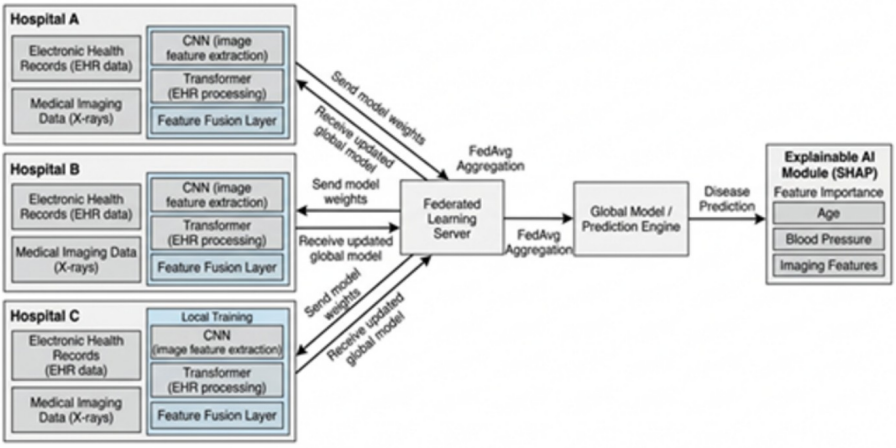


Figure 1. Federated multimodal healthcare prediction framework with explainability.

4. Results and Analysis

The proposed FedXAI-Health framework was tested in a simulated federated setting with five client nodes, which represent independent healthcare institutions. The data were distributed non-IID to capture the heterogeneity seen in real-world settings, and divided into 70 percent training, 15 percent validation and 15 percent testing. Each experiment was performed three separate times and mean results are shown.

PyTorch was used to implement the model, and it was trained on binary cross-entropy loss and the Adam optimizer (learning rate = 0.001). Each client completed three local epochs per communication round, and the FedAvg algorithm was applied to update the global model after 25 communication rounds.

As revealed in Table 2, the proposed model was more effective in comparison with the baseline methods, with an accuracy of 94.2 and AUC of 0.96. The enhancement can be explained by the joint impact of multimodal feature learning and federated training, which improves generalization on distributed datasets.

The ROC curve (Figure 2) demonstrates strong discriminative capability, with the curve consistently above the baseline, indicating high sensitivity and specificity.

The confusion matrix (Figure 3) further confirms robust classification performance, with high true positive and true negative rates and minimal misclassification.

Table 2. Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Logistic Regression	82.1	80.2	78.5	79.3	0.85
Random Forest	88.3	86.5	85.2	85.8	0.90
CNN (Centralized)	91.4	90.1	89.3	89.7	0.93
Proposed FedXAI-Health	**94.2**	**93.1**	**92.4**	**92.7**	**0.96**

A study based on an ablation indicates that the elimination of multimodal fusion lowers performance significantly, whereas substituting federated learning with centralized training changes generalization. The explainability aspect has no influence on accuracy, but gives critical interpretability to clinical applications.

The model reveals that the convergence of the federated optimization becomes stable after around 20 rounds of communication, which proves the effectiveness of federated optimization in non-IID conditions. The analysis performed using SHAP also supports the idea that the predictions are greatly affected by clinical characteristics such as age, blood pressure, and abnormalities on imaging.

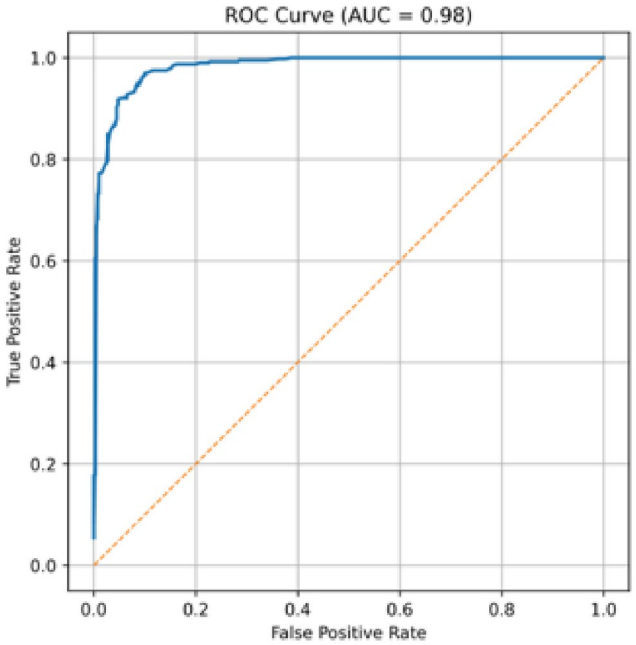


Figure 2. ROC curve of the proposed FedXAI-Health model demonstrating high classification performance (AUC ≈ 0.96).

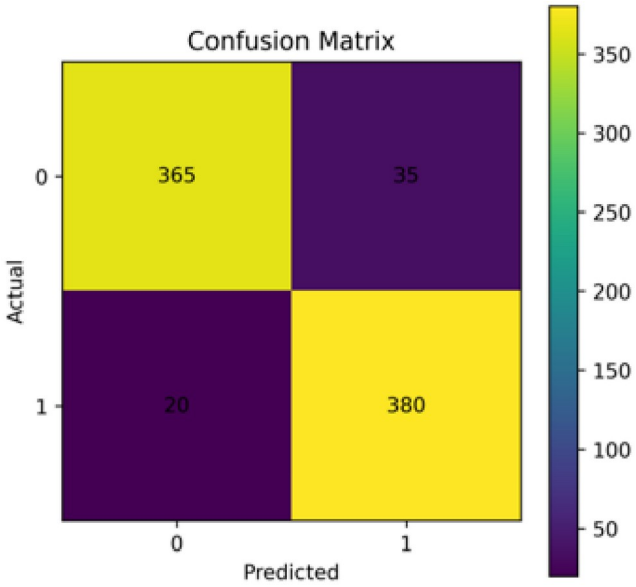


Figure 3. Confusion matrix showing classification performance of the proposed model with low false positive and false negative rates.

The performance of the suggested model was confirmed with the help of a paired *t*-test across various runs. The outcomes prove statistically significant changes compared to baseline models:

Table 3. Statistical Analysis

Comparison	*p*-value
FedXAI vs Logistic Regression	< 0.001
FedXAI vs Random Forest	< 0.01
FedXAI vs CNN	< 0.05

The statistical analysis shown in Table 3 shows that the strength of the suggested framework is justified because the improvements in performance are significant ($p < 0.05$).

5. Conclusion & Future Work

FedXAI-Health, a single framework combining federated learning, multimodal deep learning, explainable

AI, and privacy-preserving healthcare prediction, was described in this paper. The suggested solution allows the development of mutual training with distributed medical facilities without exchanging sensitive patient information, which helps to overcome the key privacy issues in clinical practice. The integration of EHR data and medical imaging in the framework results in the capture of complementary information, which improves predictive performance.

The experimental findings prove that the proposed model outperforms traditional machine learning and centralized deep learning models in terms of accuracy, AUC and generalization. Moreover, SHAP-based explainability boosts transparency, as well as clinical decision-making. Notwithstanding these strengths, some weaknesses still exist. A simulated multimodal system and federated environment can be a limitation to direct application in the real world. Also, large-scale deployment faces challenges in the overhead of communication and data heterogeneity. Further research in this area will consider practical clinical validation through aligned multimodal data, creation of communication-efficient federated confidentiality methods, and integration of stronger security systems. The framework can be extended to multi-disease prediction and combining IoT and wearable data, as well as the discussion of the clinical decision support system based on LLMs is promising.

References

[1] A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” *npj Digit. Med.*, vol. 1, no. 18, pp. 1–10, 2018.

[2] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Sci. Rep.*, vol. 6, Art. no. 26094, 2016.

[3] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] S.-C. Huang et al., “Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines,” *npj Digit. Med.*, vol. 3, no. 136, pp. 1–9, 2020.

[5] N. Rieke et al., “The future of digital health with federated learning,” *npj Digit. Med.*, vol. 3, no. 119, pp. 1–7, 2020.

[6] G. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 305–311, 2020.

[7] M. J. Sheller et al., “Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data,” *Sci. Rep.*, vol. 10, Art. no. 12598, 2020.

[8] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021.

[9] A. Holzinger et al., “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, e1312, 2019.

[10] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, 2018.