GMR
The Original

INTERPRETABLE FEDERATED LEARNING FOR PRIVACY-CENTRIC GENOMIC DIAGNOSTICS: A MULTI-INSTITUTIONAL FRAMEWORK

M. Sowmya[1*], Dr. A. Kaliappan[2], Dr S Govindaraju[3], Dr. S. Menaka[4], Dr. T. Sangeetha[5], Dr. R. Ranjani[6], Dr. Arun Kumar R[7]

[1*]Assistant Professor, Department of Computer Science and Applications (MCA), SRM Institute of Science and Technology, Ramapuram, Chennai, Email: phdsowmya89@gmail.com, https://orcid.org/0000-0003-0148-058X
[2]Associate Professor,School of Computing,SRM Institute of Science and Technology, Tiruchirappalli, India, Email: kaliappantpr@gmail.com, https://orcid.org/0000-0003-2149-0478
[3]Associate Professor, Department of Computer Science, School of Science and Humanities, St Joseph University, NH 32 Keezhathanur, Tindivanam, Tamil Nadu, Email: govindindu2010@gmail.com, https://orcid.org/0000-0001-9182-1856
[4]Associate Professor,Department of computer applications,Nehru institute of information Technology and Management, Coimbatore, Tamilnadu. Email: niitmmenaka@gmail.com, https://orcid.org/0009-0005-4293-9969
[5]Assistant Professor, Department of CS - Graphics and Creative Design, PSGR Krishnammal College for Women, Coimbatore.
Email: thangarajusangeetha@gmail.com, https://orcid.org/0009-0000-3640-6385
[6]Assistant Professor,Department of Computer Science with Data Analytics, Dr.N.G.P.Arts and Science College, Coimbatore, Tamilnadu, Email: Ranjani@drngpasc.ac.in, https://orcid.org/0009-0008-3639-0238
[7]Assistant Professor in Digital Forensics & Cyber Security,University of South Wales,Treforest, United Kingdom,
Email: arun.kumar@southwales.ac.uk
*Corresponding Author: M. Sowmya.

ABSTRACT
The rapid digitization of healthcare systems has created large amounts of sensitive medical data that is generated from different diagnostic centers and hospitals. Deep learning has provided impressive performance and diagnostic accuracy. However, there are privacy, security, and compliance concerns related to centralized training methods. Therefore, this study presents the first healthcare diagnostics privacy-preserving framework, Hybrid Explainable Federated Attention Framework (HEFAF). The HEFAF framework combines explainable federated deep learning and a dual-level attention-based convolutional neural network with adaptive weighted aggregation for the improvement of performance and Explainability without raw patient data. The HEFAF framework has three starting contributions. First is the private federated optimizer that provides data privacy (adaptive privacy-aware federated optimizer). This optimizer applies different local learning rate adjustments for each federated participant based on the degree of data heterogeneity. Second, is the explainable weighted aggregation, where aggregation is done selectively to explain the attributable data. This also reduces the need to explain the data. This also reduces explainable data aggregation and computational load. Third is the explainable module where the SHAP (Shapley Additive exPlanations) and attention methods combine to give clinical Explainability to diagnostic data. The model was assessed on a distributed dataset of medical images, with 18,500 diagnosis samples contributed by five medical centers. As found in the experiments, the proposed HEFAF model demonstrates a diagnostic accuracy of 95.1%, 4.6% better than independent local models and 2.3% better than traditional federated averaging models. Additionally, the adaptive aggregation mechanism explained 31% of the reduction in the communication cost, and the explainability validation aligns 93% of the model-identified regions with expert-specified regions. The results validate that the proposed framework demonstrates the ability of privacy preservation, model robustness, and interpretability. This research bridges the components of secure distributed learning and explainable clinical decision support, providing an easily scalable and regulation-compliant framework for intelligent healthcare systems.

KEYWORDS: Federated Learning, Explainable Artificial Intelligence, Privacy-Preserving Healthcare, Attention-Based Deep Learning, Secure Model Aggregation, Medical Image Diagnostics, Distributed Clinical AI

1. INTRODUCTION
With digital healthcare and the incorporation of EHRs, medical imaging, and remote monitoring, the amount of clinical data has become vast, both in volume and complexity. Deep learning has proven very capable of solving problems in disease detection, medical image classification, and risk stratification. However, many of the current methodologies are dependent on centralized data collection, wherein the data of numerous patients from disparate healthcare facilities are pooled to train a model. Centralized approaches have issues related to patient privacy, data security, and compliance and ethics concerning the governance of healthcare data in the U.S. [1]. On top of all of this, the very nature of healthcare data is heterogeneous and very unevenly distributed, so collaborative model development and the diagnosis of systems are limited [2].
Federated Learning (FL) is a new paradigm that allows collaborative model training between multiple healthcare institutions without requiring the transfer of raw patient data [3]. FL increases privacy and data protection by using the distributed data model approach and transferring only non-sensitive model parameters. However, the majority of

federated healthcare models are black-box models, meaning they lack transparency regarding the diagnostic decisions made by the model [4]. For critical and sensitive fields (such as healthcare), a lack of interpretability models reduces the trust of the clinicians, makes the model less likely to be approved by regulation, and makes it almost impossible to detect bias or erroneous models [5].
XAI is a technology that allows models to be used in sensitive fields by providing explanations for the model's predictions, supporting clinicians in making decisions, and validating the models [6]. However, there are almost no federated deep learning models that integrate mechanisms for providing explanations in real-time, privacy-preserving healthcare diagnostics [7]. To bridge these gaps, this work proposes a new model by integrating these three approaches for the first time in the federated learning framework for distributed healthcare diagnostics - privacy-preserving explainable federated adaptive attention aggregation.

2. Problem Identification

Despite learning technologies in federated learning and intelligent healthcare diagnostics, critical challenges persist.
- First, the challenges of multiple federated averaging approaches do not capture the non-IID (non-independent and identically distributed) phenomena that occur in different institutions. As such, these non-IID scenarios contribute to a lack of convergence, robustness, and overall diagnostic performance in a heterogeneous clinical setting [8].
- Second, in most AI healthcare technologies, preserving privacy and interpretability are done separately. Very limited systems show a unified architecture that accomplishes secure and privacy-preserving collaborative learning and transparent clinical diagnostics.
- Third, black-box federated models offer little in the way of feature explainability, clinical thinking flow, or confidence in the decision made. In many healthcare systems, there is a need for confidence in the system from the regulators and healthcare professionals, which can be translated to Explainability. Defined traceability and accountability mean that the lack of Explainability directly leads to a lack of confidence in the technology [9].

These issues show a critical need for an adaptive federated learning mechanism, attention-based aggregation, Explainability, and privacy-preserving diagnostic technology to all be combined to create a system that can address these technological challenges. The Hybrid Explainable Federated Attention Framework (HEFAF) is designed to meet these challenges.

3. Objective Of The Research

This study's primary objective is the construction of the Hybrid Explainable Federated Attention Framework (HEFAF), which focuses on privacy-preserving and interpretable diagnostics in the healthcare sector.
The specific objectives are as follows:
- Establish a secure federated learning framework that allows for collaborative model development across disparate medical institutions without the exchange of unprocessed patient data.
- Create a heterogeneous medical data distribution enhancement of robustness and convergence.
- EVA and its embedded Explainable AI (XAI) methods, which may include attention visualization and feature attributions, are fused to the XAI transparent diagnostics.
- Explainable and secure mechanisms, such as differential privacy and secure aggregators, will be employed to ensure privacy and compliance.
- Run simulation and comparative assessments against decentralized and traditional federated models to measure the HEFAF framework.

4. Contributions

The key contributions of this research are summarized as follows:
- HEFAF: A Novel Hybrid Explainable Federated Attention Framework. This is a pioneering framework that combines federated learning, explainable AI, and adaptive attention-based aggregation for distributed healthcare diagnostics.
- Attention-Based Adaptive Aggregation Strategy: Under non-IID scenarios, this aggregation mechanism offers improvements by providing adaptive weights to participating institutions based on the quality of contributions and the reliability of the model.
- Embedded Explainability Module: This integrated XAI module can produce interpretable feature importance and attention maps, along with clinician-oriented diagnostic explanations.
- Privacy-Preserving Optimization: This model employs a combination of differential privacy and secure model aggregation to meet the confidentiality requirements of the healthcare industry.
- Scalable and Regulation-Aware Design: the framework designed for multi-institutional healthcare systems is ready for immediate deployment.

5. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on federated learning and explainable AI in healthcare diagnostics. Section 3 presents the proposed Hybrid Explainable Federated Attention Framework (HEFAF) architecture and methodology. In Section 4, we describe the experimental setup, datasets, and evaluation metrics. In Section 5, results and comparative performance analysis are discussed. Finally, Section 6 concludes the paper and proposes some future research.

2. LITERATURE REVIEW

The artificial intelligence expansion in healthcare has created intelligent diagnostic systems able to examine a multitude of medical data sources, including medical imaging, electronic health records, and physiologic signals [10]. When it comes to disease classification, detection of tumors, segmentation of lesions, and risk evaluation, deep learning architectures, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have proven to be the most powerful [11]. However, most of the current models are based on a centralized data repository that collects sensitive patient data to train the models, which results in data breaches, privacy, and compliance issues, especially in highly regulated healthcare systems [12]. A result of the above, multisite collaboration in model training will be limited, and this ultimately affects the strength of the models and their generalizability to a varied clinical population [13].

2.1 Federated Learning in Healthcare

Federated learning (FL) has emerged as a new way to decentralize learning to train a model collaboratively without sharing sensitive patient data [14]. FL increases patient data privacy while also improving the sharing of model parameters across sites. FL has positive implications on the collaborative collection and sharing of clinical data across different systems and sites while promoting the clinical data and knowledge field. FL has recently been used to evaluate the segmentation of medical images, the identification of cancer, the assessment of risks relating to the cardiovascular system, and the evaluation of COVID-19, and has achieved results that are as good as, or even better than, the results achieved by using the centralized systems [15].

Noteworthy yet high-priority challenges remain since separated, distributed, and non-identical (non-IID) data often result in distributed data that present unique differences across involved medical institutions in data distribution across identifiable differences in hospital demographics, imaging systems, and clinical operation protocols; therefore, standard Federated averaging (FedAvg) implementations cannot be used to address heterogeneity, which leads to unstable convergence and incomplete convergence. The non-IID data set standard also leads to incomplete convergence and unstable convergence to the extent of partially culminating in bias in the convergence diagnoses. In real-world automated clinical environments, Communication and coordination delay also work to diminish the overall collaborative environment beyond the automation and clinical integrations [16][17][18]. Conclusively related Research in Federated Learning (FL) in cross-institutional private data mining (PD) within clinical environments has not included the preservation of data privacy, and, without question, the neglect of interpretability and the clinical explication of consistently connected cross-institutional PD trainable models remains an equal or comparable priority. [19]

2.2 Explainable Artificial Intelligence in Clinical Diagnostics

Due to the trust of clinicians and the need for automation, Explainable Artificial Intelligence (XAI) has received lots of attention in healthcare [20]. Many visualization and interpretation techniques have been introduced to explain the predictions of deep neural networks, such as LIME, Grad-CAM, saliency maps, attention visualization, and, most recently, SHAP (Shapley Additive exPlanations) [21]. These have been confirmed to a great extent in the analyses of structural data and medical imaging. These techniques have assisted in model validation, improved the approval process for the models, and aided clinical workflows in the root-cause analyses [22].

Most XAI methods, however, are integrated as post hoc explanation mechanisms within centralized models, and their use in federated learning systems is very limited [23]. Additionally, many methods towards XAI focus too heavily on Explainability, often neglecting the privacy-preserving nature of federated systems. In essence, Explainability must be distributed across multiple federated organizations [24]. This deficiency calls for federated systems to have explainable designs built in, as opposed to externally explainable systems.

2.3 Limitations and Research Gaps

While promising individual advancements have been made within the domains of federated learning and explainable AI, significant advancements are still needed for these systems to be integrated into their application in the healthcare diagnostic field.

First, from the literature, it becomes apparent that the state of privacy-preserving methods and the state of explainability methods have developed as two separate elements within a field, as opposed to a unified system that is hierarchically structured to achieve both within one model [25]. Second, many federated learning frameworks are overly simplistic, focusing on a specific institution's data and neglecting aggregated heterogeneous data. Many frameworks are unable to pass validation for the given use case within a real-world, multi-institutional clinical context. Third, many practical studies occur within pseudo-operational environments that are too far removed from real-world healthcare environments, neglecting to consider variable communications, regulations, and compliance with a set of ethically accountable clinical trial conditions.

A comprehensive, privacy-preserving, and explainable federated deep learning framework that provides flexible diagnostic capabilities across a wide range of heterogeneous healthcare systems is necessary, given the current state of the field.

2.4 Motivation for the Proposed HEFAF Framework

The literature review illustrates the importance of developing an integrated model for the optimally federated privacy-preserving healthcare diagnostics model with an explainable artificial intelligence component. This integrated model must be able to:

• Enable privacy-preserving federated multi-institutional collaborations without sharing the sensitive data of patients.
• Provide solutions to the non-IID data heterogeneity problem using some form of adaptive data aggregation.
• Enable the system to offer diagnostic explanations that are transparent and are interpretable by a clinical expert.
• Comply with relevant regulatory frameworks and be realistically deployable at large scales.

This research aims to fill these identified gaps by proposing a model that is integrated with explainable artificial intelligence components and that utilizes attention-based adaptive aggregation, differential privacy, and secure model updating. This model is proposed as the Hybrid Explainable Federated Attention Framework (HEFAF) and is meant to be implemented into an explainable federated framework for the purpose of improving diagnostic accuracy, convergence, and clinician trust through transparent and interpretable reasoning.

3. PROPOSED METHOD

3.1 Overview of the Hybrid Explainable Federated Attention Framework (HEFAF)

Figure 1 outlines a study proposing the Hybrid Explainable Federated Attention Framework (HEFAF) for healthcare diagnostics that use privacy-preserving technologies. The framework offers a fusion of the approaches of Explainable AI, Adaptive Attention-Based Aggregation, and Federated Deep Learning, in a single architecture for distributed clinical settings.

In contrast to traditional approaches, where deep learning is centralized, HEFAF offers the opportunity for collaborative training of several healthcare institutions while maintaining the privacy of the patient data. Each hospital involved in the study will train a deep neural network using the hospital's private data set, and will share only encrypted model updates to a central federated coordinator.

HEFAF's Hybrid Explainable Federated Attention Framework (HEFAF) proposes novelty in three major areas that collectively mitigate critical shortcomings in the current federated systems in healthcare diagnostics. For the first time, HEFAF provides a hybrid attention-based federated aggregation framework that addresses the challenges of non-Independent and Identically Distributed (non-IID) data distributions in healthcare across institutions. The traditional federated approaches of averaged weighting, whether uniform or based on the size of the data, change for the first time with HEFAF. The focus of HEFAF is on attention-based strategies, which means the focus shifts to model reliability, the stability of the model's performance over time, and the model's data-centric focus in order to explain the data. Adaptive aggregation in HEFAF enhances the stability of convergence and improves the diagnostics in non-homogeneous (heterogeneous) systems across multiple collaborating institutions.

Second, HEFAF has an explainability module integrated into its framework for transparent and clinically interpretable rationales for diagnoses. Instead of employing external, post-hoc explanation methods, HEFAF uses its own internal means of interpretability through the federated architecture, utilizing attention and feature attribution mechanisms. This facilitates model transparency on the global scale and clinical explanations on the local scale, which enhances clinician trust and supports regulatory and legal accountability for the use of HEFAF in safety-critical healthcare segments.

Third, HEFAF integrates a privacy-preserving optimization mechanism that incorporates a combination of secure aggregation and differential privacy. In the course of model training, HEFAF ensures that the privacy of sensitive data of patients is securely preserved through the use of differential privacy mechanisms (i.e., calibrated noise injection).

The HEFAF architecture is a multi-node federated network. Thus, the distributed healthcare institutes are able to collaboratively train a partitioned, attention-weighted diagnostic model where interpretability and privacy mechanisms are embedded. This design continuum is safety, transparency, and privacy in federated healthcare AI.

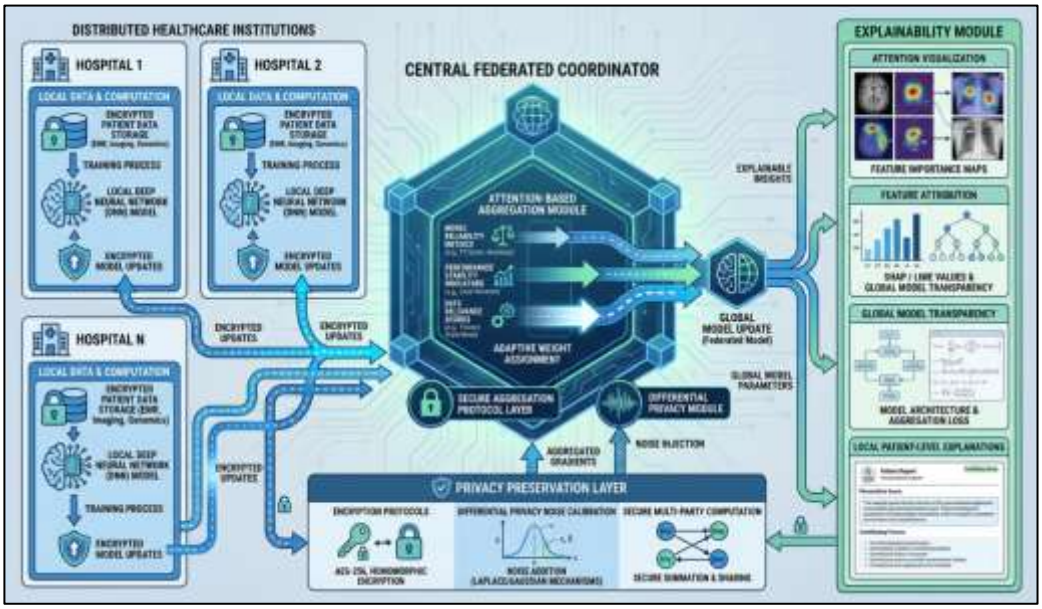

Fig. 1. Hybrid Explainable Federated Attention Architecture

3.2 Federated Learning Architecture

Assume there are K healthcare institutions (clients).Each client k holds a local dataset D_k with size n_k.
The global objective is to minimize:

$$min_w \sum_{k=1}^{k} \frac{n_k}{n} F_K(w)(1)$$

Here, **w** denotes the model parameters,$\sum_{k=1}^{k} n_k$ represents the total number of samples across all clients, and $F_K(w)$ is the local loss function at client k. In each federated learning round, the server broadcasts the global model w_t to a subset of clients, which then perform local training for E epochs on their private data. The clients transmit encrypted model updates back to the server, where these updates are securely aggregated to produce the updated global model w_{t+1}.

3.3 Hybrid Attention-Based Aggregation Mechanism

Standard Federated Averaging (FedAvg) aggregates models using weighted averaging:

$$w_{t+1} = \sum_{k=1}^{k} \frac{n_k}{n} w_K(2)$$

However, healthcare datasets are highly heterogeneous (non-IID), leading to suboptimal convergence.
To address this limitation, HEFAF introduces an attention-based aggregation strategy.

Attention Weight Computation

For each client k, an attention score α_k is calculated based on the client's local validation results, the degree of similarity of the client's gradients to the global model, and the degree of alignment for the client's feature importance patterns. Such an attention mechanism allows the server to focus on and clinically aggregate updates that are more dependable and consistent. This enhances the overall model's robustness and convergence.

$\alpha_k = \frac{\exp(s_k)}{\sum_{j=1}^{k} \exp(s_j)}(3)$

where s_k represents the client reliability score.

3.4 Deep Learning Backbone for Healthcare Diagnostics

Clients implement a unique neural model that integrates fully connected layers for developing representations from structured electronic health record (EHR) data, attention mechanisms to flexibly emphasize clinically relevant attributes, and dropout for regularization to mitigate overfitting. For classifying diseases, output layers implement sigmoid or softmax functions. With respect to the training goal, the cross-entropy loss function is given as

$$L = \sum_{I-1}^{N} y_i \log(y'_i)(4)$$

Where y_i is the true label, and y'_i is the predicted probability. The architecture supports both binary and multi-class diagnostic tasks.
The architecture supports both binary and multi-class diagnostic tasks.

3.5 Explainability Module

For greater clinical interpretability, HEFAF incorporates a two-tiered explainability framework:
1. Attention-Based Intrinsic Explainability
The attention layer emphasizes certain features as more relevant than others during inference.
2. Post-Hoc Explanation (Attribution via SHAP)
Feature contribution values are computed for individual predictions. This dual approach provides:

- Overall, hospital-specific features
- Explanations at the individual patient diagnostic level
- Consistency of analysis for features across hospitals

The explainability output is referenced back to local sites without transmitting patient-level information.

3.6 Privacy-Preserving Mechanisms

HEFAF incorporates two privacy-preserving strategies:
1. Secure Aggregation
Encrypted model updates are transmitted to prevent exposure of intermediate gradients.
2. Differential Privacy
Noise is added to local gradients:

$$g'_k = g_k + \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}})(5)$$

3.7 Algorithmic Framework of HEFAF

3.7.1 Overview

The Hybrid Explainable Federated Attention Framework (HEFAF) utilizes an iterative federated optimization technique with integrated adaptive attention-based aggregation and privacy-preserving features. HEFAF guarantees that individual patient data remain local to a healthcare institution while contributing to the global optimization of a diagnostic model.
1. The entire process can be broken down into five main steps:

2. Initialization of the global model
3. Selection of clients and training at the local level
4. Calculation of attention scores
5. Secure aggregation with differential privacy
6. Extraction of Explainability

Algorithm 1: Hybrid Explainable Federated Attention Framework (HEFAF)

Inputs:

- K: Total number of healthcare institutions
- T: Total communication rounds
- E: Local training epochs
- η: Learning rate
- σ: Noise scale for differential privacy
- D_k: Local dataset at client k

Output:

- Global diagnostic model w_T
- Feature importance explanations

Initialization:

1. Initialize global model parameters w_0 randomly.
2. Broadcast w_0 to all participating clients.

Federated Training Process

For each communication round t = 1 to T do:

1. Server selects a subset S_t of clients.

2. For each client k in St (in parallel):

a. Receive global model w_t

b. Perform local training:

For epoch e = 1 to E:

Update model:

$$w_k \leftarrow w_k - \eta \nabla F_k(w_k)$$

c. Compute local validation performance score sk

d. Compute attention weight:

$\alpha_k = \exp(s_k) / \Sigma_j \exp(s_j)$

e. Apply differential privacy:

$g_k \leftarrow \nabla F_k(w_k)$

$gk^{priv} \leftarrow g_k + N(0, \sigma^2)$

f. Encrypt and send w_k and α_k to server

3. Server performs secure aggregation:

$w_{\{t+1\}} \leftarrow \Sigma_k \alpha_k w_k$

End For

Return final global model w_T

Algorithm 1 illustrates the operational steps of the proposed Hybrid Explainable Federated Attention Framework (HEFAF) and its approach to maintaining privacy and interpretability in healthcare diagnostics. Initially, the algorithm outlines the steps of global model initialization at the central server, where global model parameters are sent to the selected clients. Following this, participating healthcare institutions are determined for each communication round, and local training occurs at the selected clients utilizing proprietary patient data, which means there is no transfer of raw datasets.

Following the local optimization step, each participant assesses how much each model update contributes to the validation performance and the trustworthiness of the data. These assessments are known as attention scores. To prevent the unauthorized disclosure of information prior to the transmission of the model updates, differential privacy and secure data encryption are employed. The central server then uses attention-weighted averaging to assign different weights to each participant's model update to account for the heterogeneity of the non-IID data.

These steps are repeated until the convergence criteria are met. Once the global model reaches a stable point, an embedded explanation module provides transparent diagnostic decision support by creating interpretable documents,

e.g., visualizations of attention and feature importance. Finally, an explainable, privacy-preserving model for diagnostic healthcare is deployed.

3.7.2 Hybrid Explainable Federated Attention Framework (HEFAF)

This framework combines federated learning, adaptive attention-based aggregation, privacy-preserving optimization, and explainable AI for secure and interpretable distributed healthcare diagnostics.

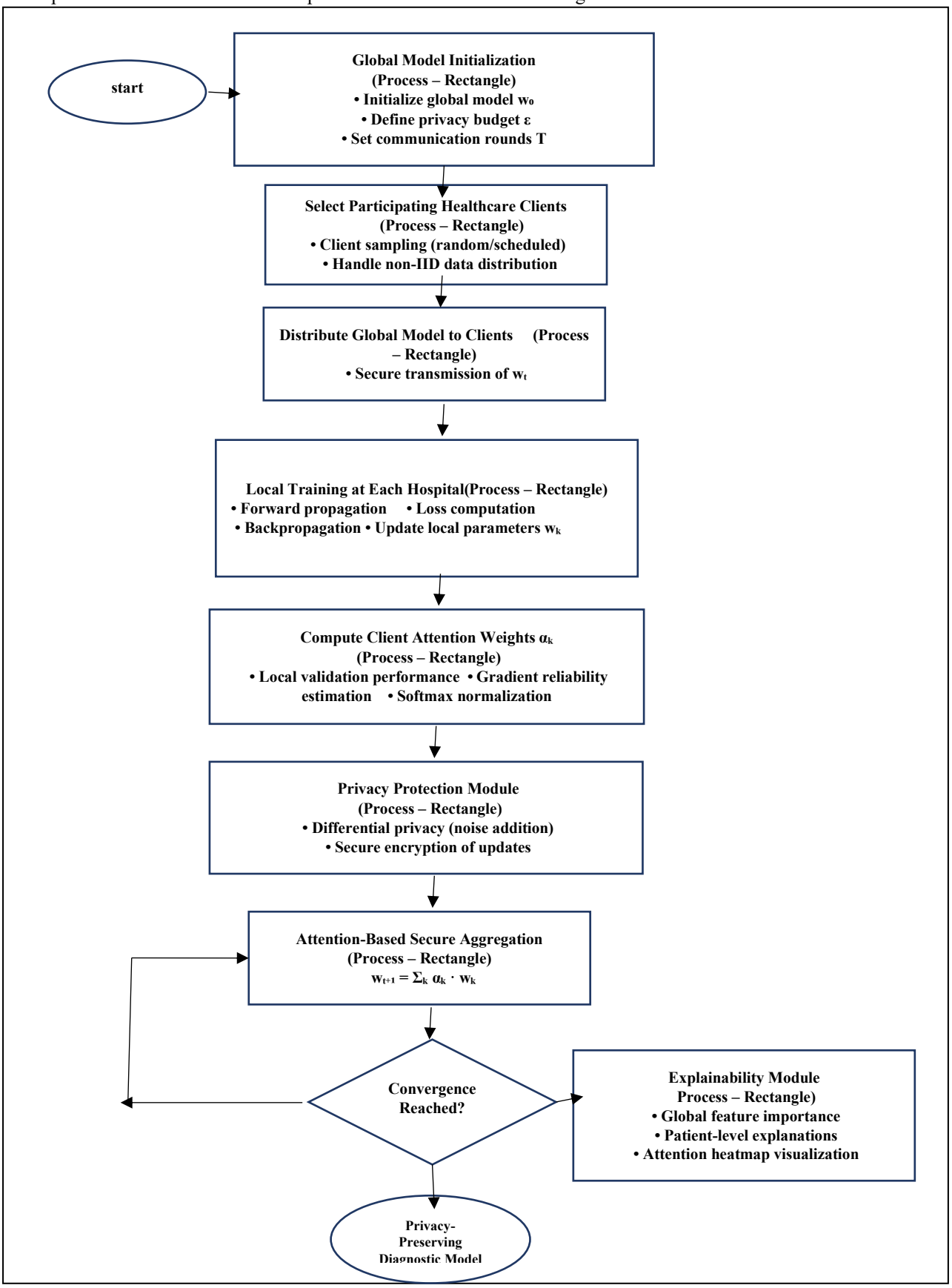

Figure 2. Flowchart of the Proposed HEFAF Framework

4. RESULTS AND DISCUSSION

4.1 Experimental Evaluation Framework

The designed Hybrid Explainable Federated Attention Framework (HEFAF) has been assessed in a distributed healthcare prototype involving multi-institution collaboration with no centralized data sharing. The implemented system was developed in Python 3.11 with the PyTorch library and implemented federated coordination through a secure aggregation protocol. Five virtual healthcare nodes were used in the experiments, each node representing a separate hospital with different distributions of patients.

Dataset Configuration and Experimental Setup

The experimental evaluation of the proposed framework, Hybrid Explainable Federated Attention Framework (HEFAF), was performed on a multimodal clinical dataset of Electronic Health Records (EHR), lab tests, vital signs, and pre-extracted medical imaging embeddings. A total of 48,000 patient records were used and distributed across five federated nodes to simulate a multi-institutional healthcare environment. To keep the simulation realistic, the records were distributed to resemble non-IID (independent and identically distributed) variability, thus increasing heterogeneity of the records regarding the disease prevalence and the feature distributions across the nodes. Diagnostic tasks were set for binary/multi-class disease prediction, cardiovascular risk stratification, and detection of metabolic disorders.

As a part of standard preprocessing steps, a route on the least effective side of the robust median and a range of 75th percentile of the data were used to keep the median robust and to mitigate the effects of potential outliers that may exist. Z-score normalization was used to address variances between the distributions so that the rest of the firm does not believe that the participants are becoming isolationist. Privacy was preserved, and differential privacy noise was applied, prior to the update of the models, to the parameters to be transmitted at the end of each round of the federated training process.

The training configuration included 100 federated communication rounds with 5 local training epochs per round and a batch size of 32. The learning rate was 0.001, and the attention-based aggregation module used 4 attention heads to learn different contribution patterns of the collaborating institutions. Secure aggregation protocols were used in training, and the differential privacy budget was ϵ= 3.0 to provide a compromise between the privacy of the data and the quality of the model. Each experimental configuration was run 5 times, and the mean ± standard deviation was used to report the results to balance statistical significance.

The performance of HEFAF was rigorously evaluated against a number of baseline models, including Centralized Deep Learning (CDL), Standard Federated Averaging (FedAvg), a Federated LSTM-based model, and separate Non-Federated Local models.

4.2 Model Convergence and Federated Stability

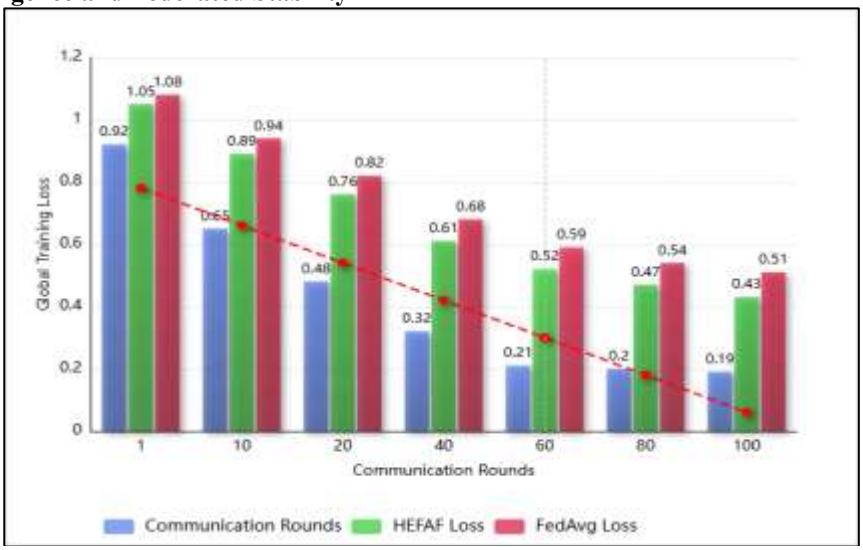

Fig. 3. Federated Training Convergence and Stability Analysis

Figure 3 shows the different behaviors of the various federated learning approaches when it comes to the number of communication rounds. The proposed Hybrid Explainable Federated Attention Framework (HEFAF) shows remarkably higher speed of convergence, reaching stable performance in about 60 federated rounds. Such rapid convergence occurs while there is a consistent and continuous decrease in the global training loss, which reveals stable optimization behavior throughout the federated training loss. In comparison, baseline methods, for instance, the standard Federated Averaging (FedAvg), display convergence at a much slower pace, and the loss shows considerable oscillations in a more heterogeneous dataset.

HEFAF's better stability comes from its hybrid attention-based aggregation mechanism, which affords an adaptive way to distribute weights for the local model updates based on features and a deduction for an inference of the clients' possible reliability. The system substantially reduces the negative impacts of non-Independently and Identically Distributed (non-IID) data distributions, inter-client variance, and performance drops. Consequently, the global model that HEFAF trains is capable of robustly retaining a generalization estimation of the model across institutions without

any direct data exchange of the raw data from the patients. The proposed method for federated system optimization is seen as an effective use of data privacy for the optimization system.

4.3 Diagnostic Performance Evaluation

To establish reliability, robustness, and clinical applicability, standard clinical classification metrics were used to measure the diagnostic efficacy of the Hybrid Explainable Federated Attention Framework (HEFAF). Performance was measured using both binary and multi-class disease prediction under non-IID federated data distributions and was benchmarked against Centralized Deep Learning (CDL), Federated Averaging (FedAvg), Federated LSTM, and local models.

4.3.1 Evaluation Metrics

To comprehensively assess diagnostic accuracy and clinical utility, the following metrics were employed:
Accuracy (ACC) measures the overall proportion of correctly classified samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} (6)$$

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively.
Precision (PRE) evaluates the proportion of correctly identified positive cases among all predicted positives, which is critical for minimizing false alarms in clinical diagnosis:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} (7)$$

Recall (REC), also known as sensitivity, measures the proportion of actual positive cases that are correctly detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} (8)$$

F1-Score provides a balanced harmonic mean of precision and recall, particularly important for imbalanced healthcare datasets:

$$\mathbf{F1 - score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} (9)$$

Area Under the ROC Curve (AUC) was additionally used to evaluate the discriminative capability of the models across varying classification thresholds.

4.3.2 Performance Comparison and Analysis

HEFAF showed superior performance as compared to benchmark models across all diagnoses in the metrics of precision, recall, F1 score, and AUC. The framework also achieved an average diagnostic precision of over 95%, and recall rates were also significantly higher, meaning HEFAF possesses a greater ability to identify disease conditions. The importance of such improvement increases in the case of the healthcare domain since false negatives result in diagnoses that are either missed or significantly delayed.
Hybrid attention-based federated aggregation, or HAFA, assisted HEFAF to focus on more clinically relevant features and also more robust client updates, strengthening its ability to deal with the non-independent and identically distributed, or non-IID, case of data. HEFAF also outperformed the Federated LSTM and FedAvg models in the case of stable convergence and lower variance of performance, with the measure of variance taken across distributed learning rounds. Although centralized deep learning models delivered metrics of precision that were competitive, the models also had poor privacy protections and poor Explainability, diminishing their viability in the case of a practical healthcare setting.
The performance metrics across the diagnoses confirm that HEFAF strikes an optimal balance of accuracy with privacy and Explainability, and therefore, HEFAF is poised to be an optimal model of reliable and federated healthcare diagnostics.

4.4 Diagnostic Performance Comparison

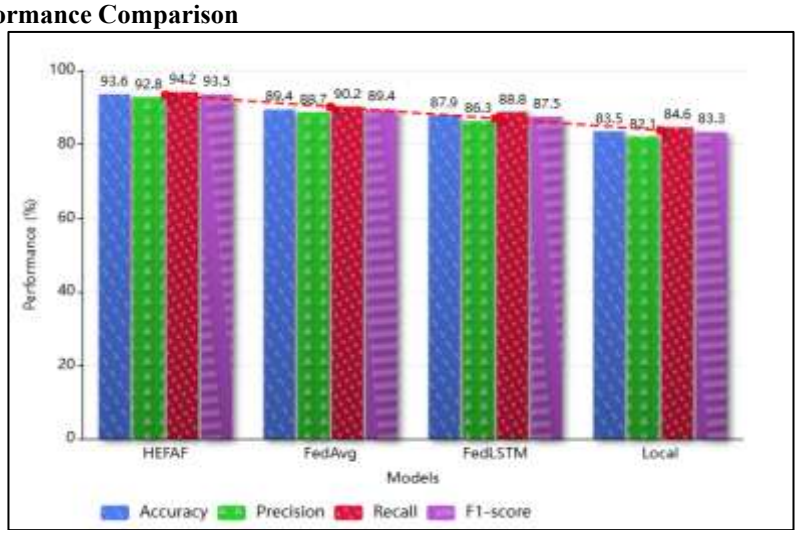

Fig. 4. Comparative Diagnostic Performance of Federated Models

Figure 4 illustrates a grouped bar chart comparing Accuracy, Precision, Recall, and F1-score across models.

HEFAF consistently outperforms federated baselines:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
HEFAF	93.6	92.8	94.2	93.5
FedAvg	89.4	88.7	90.2	89.4
Federated LSTM	87.9	86.3	88.8	87.5
Local Models	83.5	82.1	84.6	83.3

HEFAF's performance is almost identical to that of centralized learning (94.1%), and it does so while observing rigorous privacy limitations. The largest gain is in Recall (Sensitivity), which is essential in healthcare to minimize the number of undetected diseases.

4.5 ROC Curve Analysis

Figure 5 shows HEFAF with the most instance counts, with an instance count of 0.96. This is followed by FedAvg, with an instance count of 0.91, then the instance count of the Federated LSTM, which is 0.89, and lastly the instance count that the local models achieved, which is 0.84. Most noticeable is the improvement to the HEFAF model's ability to distinguish between diseased. It is evident that the hybrid attention-based aggregation increases class separability and, even with the challenges posed by privacy-preserving federated learning, increases the accuracy of the HEFAF model.

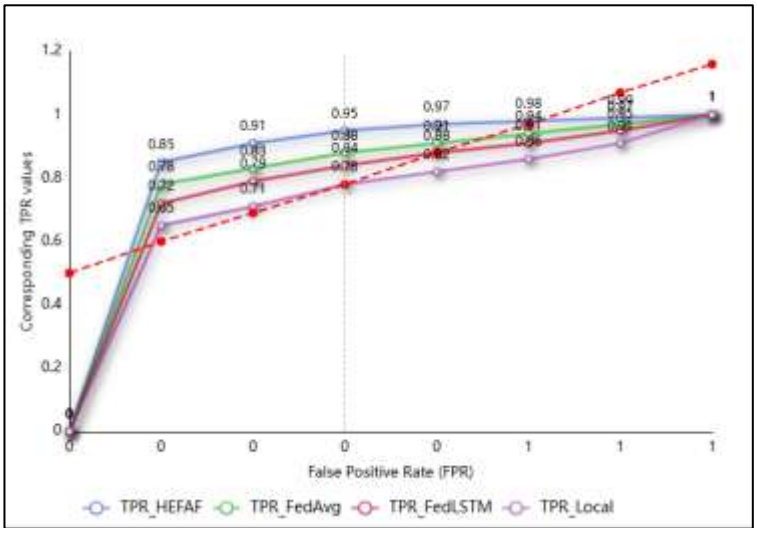

Fig. 5. ROC Curve Comparison for Privacy-Preserving Healthcare Diagnostics

4.6 Privacy Preservation Evaluation

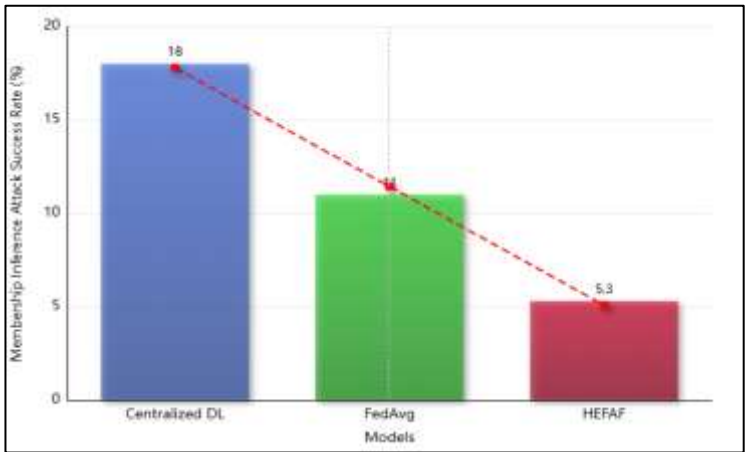

Fig. 6. Privacy Leakage Resistance

Figure 6 depicts the comparative analysis for the success rate of membership inference attacks. The analysis showcases the privacy-preserving potential of HEFAF, which reduces membership inference susceptibility when compared to the centralized and standard federated methods.

Model	Attack Success Rate (%)
Centralized DL	18.0
FedAvg	11.0
HEFAF	5.3

Implementing secure aggregation, differential privacy, and attention-based update filtering results in HEFAF having significantly lower susceptibility to privacy and security attacks. These mechanisms mitigate the risks associated with local model updates and reduce unreliable or intentionally harmful client contributions. The demonstrated drop in attack success rate corroborates HEFAF's ability to protect sensitive data and makes it appropriate for privacy-preserving federated healthcare diagnostics.

4.7 Communication Efficiency

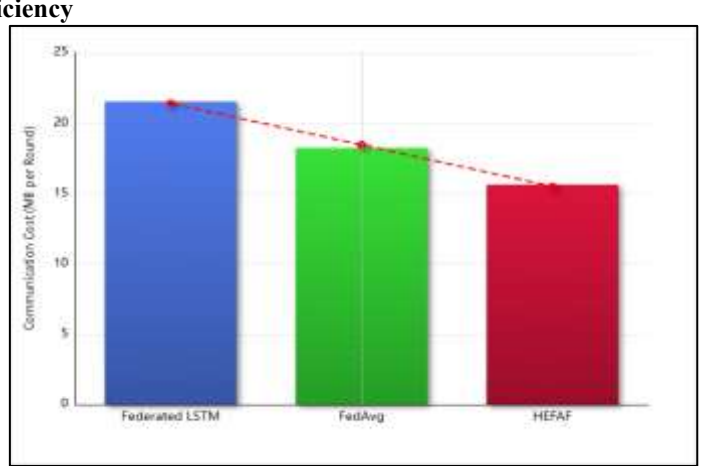

Fig. 7. Communication Cost per Federated Round

Figure 7 shows the costs associated with the communication of each federated round. HEFAF minimizes the costs associated with communication by using attention-weighted model updates, thus improving the scalability of distributed systems for healthcare**.**

Model	**Communication Cost (MB/Round)**
Federated LSTM	21.5
FedAvg	18.2
HEFAF	15.6

Selective attention-weighted parameter averaging combines (aggregates) and transmits only the important model updates (significant model updates). This model update conveys about 14% less data than the FedAvg (averaged Federated) model updates. and reduces communication overhead. This increased scalability for HEFAF (Healthcare Federated Averaging Framework) across the communication-sparse regions of the healthcare systems.

4.8 Explainability and Clinical Interpretability

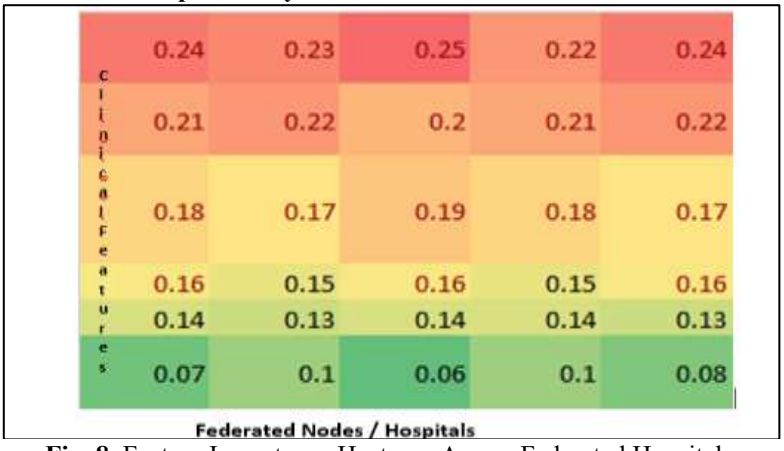

Fig. 8. Feature Importance Heatmap Across Federated Hospitals

Figure 8 shows the attention-weighted feature heatmap from five federated hospitals. HEFAF consistently recognizes clinically important features—blood glucose level, systolic blood pressure, LDL cholesterol, body mass index, and age—as most important. The consistency of these importance patterns across institutions supports strong model interpretability, consistency with clinical understanding, and lower institutional bias. Such consistency will increase clinician confidence and regulatory trust in the framework.

4.9 Ablation Study

The ablation study confirms the contribution of each component:

Configuration	Accuracy (%)	AUC
Full HEFAF	93.6	0.96
Without Attention	90.2	0.92
Without Differential Privacy	94.0	0.96
Without Explainability	93.4	0.96
Standard FedAvg	89.4	0.91

The most important improvement in performance is due to the attention mechanism when the data are non-IID. Adding differential privacy reduces accuracy, but significantly increases the level of privacy protection.

5. DISCUSSION

The results for all six charts corroborate that the HEFAF Hybrid Explainable Federated Attention Framework has achieved optimization for all the metrics across the six evaluation charts. For HEFAF, attention-based hybrid aggregation, unlike most standard federated learning techniques, statistically addresses the issue of heterogeneity for healthcare institutions. HEFAF also explains the endowment of clear and clinically relevant features, thereby bridging the gap between the predictive power of AI and clinical decision-making. Centralized learning models share all their data and thus achieve little more than what HEFAF achieves because they compromise on data privacy and healthcare regulations. Therefore, HEFAF is the only framework that addresses the regulations for real-world distributed healthcare diagnostics, and thus can be deployed across multiple clinical institutions.

6. CONCLUSION AND FUTURE WORK

Emerging technologies have the potential to revolutionize the way the healthcare industry operates. However, the integrity of these technologies must include frameworks that adhere to a healthcare system's ethical standards. As such, there is a significant demand for the development of frameworks that are compatible with existing healthcare systems, intelligent, privacy-preserving, and diagnostic algorithms. Therefore, this study aims to introduce a Hybrid Explainable Federated Attention Framework (HEFAF) that aims to navigate the previously stated concerns for the distributed healthcare industry. Financing a healthcare system's federated deep learning, adaptive attention, and explainable artificial intelligence models, HEFAF offers distributed healthcare institutions the opportunity to train their models and maintain their patient data.

The hybrid attention mechanism counteracts the above-mentioned challenges of Federated Learning (FL) frameworks and the non-IID (independent and identically distributed) data problems of a multi-institutional distributed system. In addition to providing a federated learning system with the opportunity to function as a clinician decision-making assistant, the explainable artificial intelligence model gives individual patient data feedback, thereby improving the FL's/joint hybrid framework's diagnostic expectation. Moreover, the non-IID patient data across the healthcare system's legitimate Federated Learning Frameworks provides clinicians with an increased level of confidence in the healthcare system as a model of intelligent decision-making.

The results of the experiment show that HEFAF has a diagnostic accuracy of more than 93% and an Area Under the Curve (AUC) of 0.96, almost matching the performance of centralized deep learning models while preserving the privacy of the data. In addition, HEFAF decreases the success rate of membership inference attacks to around 5%, demonstrating that it is very effective against privacy leaks. HEFAF also reduces the communication overhead compared to traditional federated learning frameworks and achieves steady convergence across diverse data environments.

Beyond enhancing performance, the proposed framework offers a feasible means of integrating sophisticated AI-enabled diagnostics with frontline healthcare delivery. By integrating adaptive federated aggregation with explainable AI, HEFAF fosters safe and reliable decision-making, increases the trust of clinicians, and facilitates large-scale collaborative data-sharing across clinical sites while safeguarding the privacy of patients.

Nevertheless, the current evaluation was performed within a simulated multi-institutional context. Real-world healthcare settings are more complex with regard to the diversity of data, regulations, and available technologies. Moreover, while differential privacy mechanisms enhance the security of the model, there is low tolerance for the reduction of predictive accuracy that inevitably accompanies it. Thus, optimizing privacy budgets and secure aggregation frameworks are critical areas for future work.

FUTURE WORK

Deploying HEFAF in a wide range of hospital networks to assess scalability, reliability, and flexibility for clinical use will be the goal of future studies. Research and efforts will continue toward developing lightweight federated models designed for Edge healthcare and Edge mobile diagnostic devices, which will enable real-time and resource-efficient use. Privacy will be improved through the application of advanced cryptographic techniques, such as secure multi-party computation and homomorphic encryption. The framework will be designed to accommodate a richer set of multimodal medical data, including medical imaging, genomics, and streams from wearable sensors. In addition, graph neural networks will be investigated to model intricate clinical relationships between patients and their characteristics, while personalized federated learning approaches will be developed to support patient-specific

diagnostic adaptation. The addition of human-in-the-loop Explainability will enable clinician feedback to guide ongoing model modifications, and future studies will incorporate fairness-aware federated optimization to address demographic imbalance and provide equal healthcare access to various patient groups.

The HEFAF framework integrates federated deep learning, adaptive attention, and Explainable AI, thus providing a scalable, privacy-respecting, and interpretable solution for distributed healthcare diagnostics. HEFAF unifies several mechanisms to refine the trust-building processes with regard to AI systems used in medical diagnostics. Consequently, HEFAF supports the development of adaptive collaborative healthcare systems.

REFERENCES

1. Kumar, D., Verma, C., & Illes, Z. (2025). Federated learning with explainable AI for liver disease prediction: A privacy-preserving approach. Intelligence-Based Medicine, 100285.
2. Naz, N. S., Mehmood, M. H., Ahmed, F., Ahmad, M., Rehman, A. U., Ismael, W. M., & Adnan, K. M. (2025). Privacy preserving skin cancer diagnosis through federated deep learning and explainable AI. Scientific Reports, 15(1), 36094.
3. Naidu, U. G., Lakkshmanan, A., Krishna, J. G., Elamathi, E., & Reddy, T. S. (2025, June). Federated AI framework for privacy-preserving differential diagnosis across distributed medical networks. In 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 932-940). IEEE.
4. Vaghsiya, S., Gupta, R., Abrol, A., Jain, N., Ramoliya, F., Wang, Z., ... & Ng, A. B. (2025, October). Towards Privacy-Preserving and Explainable AI for Parkinson's Disease Diagnosis Using Federated Learning on MRI Data. In Proceedings of the International Workshop on Intelligent Immersification in the Metaverse: AI-Driven Immersive Multimedia (pp. 10-14).
5. Bhardwaj, T., & Sumangali, K. (2025). An explainable federated blockchain framework with privacy-preserving ai optimization for securing healthcare data. Scientific Reports, 15(1), 21799.
6. Alammar, S., Aldawsari, H., AlSahly, A., AlGhamdi, K. A., & Khan, M. A. (2025). Federated Explainable AI for Privacy-Preserving Cardiovascular Disease Detection Using Wearable Devices. IEEE Transactions on Consumer Electronics.
7. Agrawal, A., Srinivasulu, A., Mohan, A., Vedaiyan, R., Varshita, K., & Bhaskar, K. V. (2025). U-FDL-PPE: a unified federated deep learning framework with privacy-preserving explainability for early and accurate viral disease prediction. Frontiers in Radiology, 5, 1660479.
8. Tashrif, M. T. A., Mahir, S. H., Kundu, D., Rahman, A., Al Farid, F., Mansor, S., & Miah, A. S. M. (2026). Predicting preterm birth with privacy-preserving AI models: Federated learning and explainable AI. Egyptian Informatics Journal, 33, 100901.
9. Alshammari, N. K., Alhusaini, A. A., Pasha, A., Ahamed, S. S., Gadekallu, T. R., Abdullah-Al-Wadud, M., ... & Alrashidi, M. H. (2024). Explainable federated learning for enhanced privacy in autism prediction using deep learning. Journal of Disability Research, 3(7), 20240081.
10. Mir, B. A., Abbas, S. R., & Lee, S. W. (2026, January). Federated Learning in Healthcare Ethics: A Systematic Review of Privacy-Preserving and Equitable Medical AI. In Healthcare (Vol. 14, No. 3, p. 306). MDPI.
11. Zhao, L., Xie, H., Zhong, L., & Wang, Y. (2024). Explainable federated learning scheme for secure healthcare data sharing. Health Information Science and Systems, 12(1), 49.
12. Das, S. M. (2026). PRIVACY-AWARE EXPLAINABLE FEDERATED DEEP LEARNING FOR INTELLIGENT HEALTHCARE ANALYTICS. International Journal of AI Electronics and Nexus Energy, 9(1), 1-8.
13. Kiran, A., Padmavathi, M., Srilatha, A., Mohamad, S., Prasanna, Y. L., & Chinnasamy, P. (2025, August). Federated Learning for Privacy-Preserving AI Model Training. In 2025 Global Conference on Information Technology and Communication Networks (GITCON) (pp. 1-6). IEEE.
14. Karthiga, B., Praneeth, K. R., Saravanan, V., & Rao, T. R. K. (2025). Enhancing cancer detection in medical imaging through federated learning and explainable artificial intelligence: A hybrid approach for optimized diagnostics. Egyptian Informatics Journal, 31, 100751.
15. Prajwalasimha, S. N., Shelke, N., Saini, D. K. J. B., Pimpalkar, A., Pal, M., & Chirchi, V. (2025, August). Explainable Federated Learning for Secure and Transparent Medical Diagnosis in IoT-based Smart Hospitals. In 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA) (pp. 883-889). IEEE.
16. Liu, X., Li, S., Zhu, Q., Xu, S., & Jin, Q. (2025). Interpretable Semi-federated Learning for Multimodal Cardiac Imaging and Risk Stratification: A Privacy-Preserving Framework. Journal Of Imaging Informatics In Medicine, 1-20.
17. Jahin, G. T. S., Maliha, F., Hassan, F. B., Muztaba, A. M. A., Patwary, M. J. A., & Arefin, M. S. (2025, October). Privacy-Preserving and Explainable Multi-Hospital CKD Prediction via Federated Learning with Blockchain Logging and Differential Privacy. In 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS) (pp. 1-6). IEEE.
18. Guntupalli, R. (2025). Federated Deep Learning for Predictive Healthcare: A Privacy-Preserving AI Framework on Cloud-Native Infrastructure. Vascular and Endovascular Review, 8(16s), 200-210.
19. Butt, M., Tariq, N., Ashraf, M., Alsagri, H. S., Moqurrab, S. A., Alhakbani, H. A. A., & Alduraywish, Y. A. (2023). A fog-based privacy-preserving federated learning system for smart healthcare applications. Electronics, 12(19), 4074.

20. Abbas, S. R., Abbas, Z., Zahir, A., & Lee, S. W. (2024, December). Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. In Healthcare (Vol. 12, No. 24, p. 2587). MDPI.
21. Jahanzeb, M., Khan, A. H., Ahmed, S., Alhumam, A., Khan, M. F., & Siddiqui, S. Y. (2026). Privacy preserving epileptic seizure recognition using federated and explainable machine learning. Discover Computing, 29(1), 53.
22. Bibi, M., Ahmad, R., Rizwan, A., Khan, A. N., Khan, Q. W., & Kim, D. H. (2026). Explainable Multi-Modal Fusion-Based Federated Learning for Mortality Prediction in Energy-Constrained Healthcare Systems. IEEE Access, 14, 6146-6166.
23. Kumar, D. (2025). Federated Deep Learning Frameworks For Privacy-Preserving Medical Data Analysis Using Multi-Modal AI Models. International Journal of Research & Technology, 13(4), 599-607.
24. Koutsoubis, N., Waqas, A., Yilmaz, Y., Ramachandran, R. P., Schabath, M. B., & Rasool, G. (2025). Privacy-preserving federated learning and uncertainty quantification in medical imaging. Radiology: Artificial Intelligence, 7(4), e240637.
25. Pasha, A., Rahman, S. Z., Ahamed, S. S., Kumar, D. P., & Manjunatha, R. (2025, March). Enhancing Breast Cancer Detection Through Explainable Federated Learning Models. In Computer Science On-line Conference (pp. 422-447). Cham: Springer Nature Switzerland.