International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(12s) 2026 pp. 330-353
https://doi.org/10.70917/ijcisim-2026-3898
Article

FEMT-GAT: An Explainable Federated Multimodal Transformer–Graph Attention Network for Privacy-Preserving Disease Prediction

A. Raghavendra Rao[1], C. K. Gomathy[2]

[1]Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV), Enathur, Kanchipuram-631561, TamilNadu, India.
Email: raghavamay15@gmail.com
[2]Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV), Enathur, Kanchipuram-631561, TamilNadu, India.
Email: ckgomathy@kanchiuniv.ac.in

Abstract: The growing use of multimodal clinical data creates an opportunity to improve disease prediction, but conventional centralized learning requires hospitals to pool sensitive patient records and often provides limited interpretability. This paper presents FEMT-GAT, an explainable Federated Multimodal Transformer–Graph Attention Network for privacy-preserving disease prediction across geographically distributed healthcare institutions. The framework uses modality-specific encoders for electronic health records, laboratory measurements, demographic variables, and medical images. A cross-modal transformer learns context-dependent interactions among available modalities, while a patient similarity graph and multi-head graph attention network incorporate relational evidence from clinically comparable cases. Collaborative training is performed through federated optimization so that raw data remain at the originating institutions. Update clipping, differential privacy, encryption, and secure aggregation are incorporated to reduce information leakage. The inference module produces disease probabilities together with modality-level, feature-level, image-level, and graph-level explanations using attention rollout, modality ablation, SHAP or integrated gradients, Grad-CAM, and graph attention coefficients. The proposed architecture supports incomplete modalities through masking and modality dropout and can be adapted to binary diagnosis, multiclass classification, risk regression, and survival analysis. FEMT-GAT therefore provides a unified methodology for accurate, transparent, and privacy-aware collaborative clinical intelligence.

Keywords: —Federated learning, multimodal learning, transformer, graph attention network, explainable artificial intelligence, differential privacy, secure aggregation, disease prediction.

1. Introduction

Artificial intelligence has become an important enabling technology for computer-aided diagnosis, clinical risk stratification, disease screening, prognosis estimation, and personalized treatment planning. Modern hospitals routinely generate heterogeneous information, including electronic health records, laboratory measurements, physiological signals, demographic attributes, clinical notes, radiographs, computed tomography scans, magnetic resonance images, ultrasound images, and histopathology slides. Each modality describes a different aspect of a patient's health condition. Medical images capture anatomical or pathological structures, laboratory tests quantify biochemical and physiological abnormalities, clinical narratives describe symptoms and medical history, and demographic variables provide contextual risk information. Therefore, reliable disease prediction generally requires joint interpretation of complementary evidence rather than dependence on a single data source.

Conventional machine-learning systems often use manually selected clinical variables and relatively shallow classifiers. Although such methods can be useful for small structured datasets, they may fail to capture complex nonlinear relationships among variables and cannot directly process high-dimensional unstructured inputs such as medical images or free-text clinical notes. Deep neural networks have improved representation learning by automatically extracting discriminative features from raw data. However, many existing clinical deep-learning models are developed for only one modality, such as image-only disease classification or structured-record-based risk prediction. This unimodal design restricts the model's ability to reproduce the multimodal reasoning process followed by clinicians.

The transformer architecture introduced the self-attention mechanism as a powerful alternative to recurrent and convolutional sequence models [1]. Self-attention enables a model to estimate the contextual importance of each token relative to all other tokens and supports parallel computation. These properties have led to transformer applications in natural-language processing, medical imaging, electronic health records, physiological signals, and multimodal clinical prediction. A unified multimodal diagnostic transformer can jointly process medical images, clinical complaints, laboratory tests, and demographic attributes, allowing clinically related observations from different sources to reinforce one another [2]. Nevertheless, transformer-based multimodal models usually treat patients as independent samples and do not explicitly exploit relationships among clinically similar patients.

Graph neural networks provide a complementary mechanism for modelling relationships. In a patient graph, each node represents an individual patient, while edges represent similarity in symptoms, imaging patterns, laboratory profiles, disease stage, or other clinically relevant features. Graph Attention Networks assign adaptive weights to neighbouring nodes and therefore learn which related patients provide useful evidence for the target prediction [3]. This approach is particularly valuable for rare diseases, ambiguous cases, imbalanced datasets, and patient subgroups with limited local samples.

Despite the potential of multimodal and graph-based learning, healthcare data cannot usually be pooled into a single central repository. Patient records are protected by legal, ethical, institutional, and security requirements. Hospitals may also use different data-management systems and may be unwilling to transfer sensitive information to an external server. Federated learning addresses this limitation by allowing institutions to train a common model while keeping raw data within each hospital. In the original federated averaging approach, clients perform local optimization and transmit model updates to a coordinating server, which aggregates them into a new global model [4]. Medical studies have shown that federated learning can facilitate multi-institutional collaboration without directly sharing patient records [5], [6].

However, federated learning alone does not guarantee complete privacy. Model updates can potentially reveal information about local training data. Differential privacy limits the influence of any single record by clipping updates and injecting calibrated noise, while secure aggregation prevents the coordinating server from inspecting individual client updates. The combination of federated optimization, differential privacy, and secure aggregation is therefore important for privacy-preserving medical artificial intelligence.

Another major requirement is explainability. Clinical users must understand why a model produces a particular diagnosis or risk score. Black-box predictions may be difficult to validate, may conceal spurious correlations, and may reduce trust among physicians and patients. Explainability techniques such as SHAP, integrated gradients, Grad-CAM, transformer attention rollout, and graph-neighbour attribution can identify influential clinical variables, image regions, modalities, and related patients. Explainability must be integrated carefully because attention values alone do not always provide a complete causal explanation. A clinically useful system should combine multiple complementary explanation levels and present the results in a concise human-readable form.

The proposed FEMT-GAT framework is motivated by these interconnected challenges. It integrates modality-specific feature extraction, multimodal transformer fusion, graph attention-based patient relationship modelling, federated learning, differential privacy, secure aggregation, and explainable artificial intelligence in a unified architecture. The goal is to improve disease-prediction performance while preserving institutional data ownership and providing transparent evidence for the resulting clinical decisions.

Healthcare institutions possess large quantities of valuable clinical information, but the information is fragmented across different hospitals, modalities, acquisition protocols, and data-management systems. Centralized deep-learning approaches require raw data to be transferred to a common server. Such transfer increases privacy and security risks and may conflict with institutional policies and data-protection requirements. As a result, individual

hospitals often train models using only their local data, which may be limited, imbalanced, or biased toward the population served by that institution.

Existing unimodal models fail to exploit complementary interactions among medical images, laboratory measurements, symptoms, clinical history, and demographic information. Simple multimodal feature concatenation does not adequately model contextual relationships and may allow high-dimensional modalities to dominate the decision. Transformer-based fusion improves multimodal representation learning but commonly treats patients independently. Graph-based methods model patient relationships but may not effectively integrate heterogeneous clinical modalities. Federated learning supports distributed collaboration, yet conventional federated averaging may be unstable when hospital datasets are non-identically distributed. Furthermore, many federated diagnostic systems provide limited interpretability and do not simultaneously explain modality importance, image regions, clinical variables, and influential graph neighbours.

Therefore, there is a need for an integrated disease-prediction framework that can: (i) learn complementary representations from heterogeneous clinical modalities; (ii) capture relationships among similar patients; (iii) support collaborative training across hospitals without sharing raw records; (iv) protect local updates against privacy leakage; (v) remain robust under missing modalities and non-IID institutional data; and (vi) generate clinically meaningful explanations for every prediction.

The primary aim of this work is to develop an explainable and privacy-preserving federated multimodal learning framework for collaborative disease prediction. The proposed FEMT-GAT model combines transformer-based multimodal fusion and graph attention-based patient modelling within a secure federated-learning environment. It is designed to produce accurate predictions from distributed clinical datasets while ensuring that patient records and local graph structures remain within the originating healthcare institution.

The major objectives of the proposed research are:

1. To design modality-specific preprocessing and encoding modules for structured clinical records, medical images, laboratory measurements, demographic variables, and clinical narratives.

2. To project heterogeneous modality representations into a common latent space and perform context-aware fusion using a multimodal transformer.

3. To construct privacy-preserving local patient-similarity graphs and learn relational information using multi-head graph attention.

4. To enable collaborative model training across multiple hospitals through federated averaging or heterogeneity-aware federated optimization.

5. To apply update clipping, differential privacy, encryption, and secure aggregation to reduce the risk of patient-information leakage.

6. To support incomplete multimodal records through modality masking and modality-dropout training.

7. To generate multimodal, feature-level, image-level, and graph-level explanations for disease predictions.

8. To evaluate the framework using predictive performance, robustness, privacy, convergence, communication cost, and interpretability measures.

The principal contributions of FEMT-GAT are summarized as follows:

• A unified architecture is proposed for combining multimodal transformer fusion, graph attention learning, federated optimization, and explainable artificial intelligence.

• The model represents each patient using complementary information from heterogeneous clinical modalities instead of relying on a single source.

• A local patient graph is constructed at every participating hospital, ensuring that relational structures and patient identifiers are not transferred to the central server.

• Graph attention assigns adaptive importance to clinically similar patients and enhances the representation of rare, ambiguous, or underrepresented cases.

• Privacy-preserving federated training is achieved through update clipping, differential privacy, secure aggregation, and optional heterogeneity-aware optimization.

• The explainability module combines modality ablation, clinical-feature attribution, medical-image localization, and graph-neighbour importance to produce a multi-level clinical explanation.

• The modular design supports binary classification, multi-class diagnosis, clinical-risk regression, and other healthcare prediction tasks.

The proposed framework is intended for distributed healthcare environments in which multiple institutions possess related but non-identically distributed datasets. The architecture can process structured and unstructured clinical information and can be adapted to different medical specialties by replacing the modality encoders and modifying the patient-similarity criteria. The work focuses on disease prediction and clinical-risk assessment rather than autonomous treatment recommendation. The model is intended to support qualified healthcare professionals and should not replace clinical judgment.

The framework assumes that each participating institution has sufficient computational resources to perform local training and that the institutions agree on a common model architecture, feature definitions, label space, and communication protocol. The central server coordinates optimization but does not receive raw patient data, local embeddings, graph edges, or direct identifiers. The framework may be evaluated using simulated federated partitions or real multi-institutional datasets, depending on data availability and ethical approval.

The remainder of the work is organized as follows. Chapter 3 presents the related work on multimodal clinical learning, transformer-based diagnosis, graph neural networks, federated learning, privacy-preserving optimization, and explainable artificial intelligence. Chapter 4 describes the proposed FEMT-GAT methodology, while the subsequent chapter presents the experimental results, discussion, conclusions, and future research directions.

2. Related Work

Machine-learning and deep-learning techniques have been widely investigated for disease classification, prognosis, mortality prediction, and clinical decision support. Early studies typically relied on structured variables extracted from electronic health records and used logistic regression, decision trees, support vector machines, random forests, or shallow neural networks. These approaches offered interpretable relationships between a limited number of variables but required manual feature engineering and often failed to capture complex temporal and nonlinear interactions.

Deep convolutional neural networks significantly improved automated analysis of radiographs, CT scans, MRI scans, and histopathology images. Nevertheless, image-only models may produce incomplete decisions because visually similar abnormalities can correspond to different diseases, while some clinically important conditions may have subtle imaging signs. Similarly, models based only on structured EHR data cannot directly use detailed anatomical information. These limitations encouraged the development of multimodal systems that combine imaging, text, laboratory values, and demographic information.

The transformer architecture proposed by Vaswani et al. uses self-attention to model long-range contextual dependencies without recurrent operations [1]. Its modality-independent attention blocks have enabled the same architectural principles to be used for text, images, time series, and multimodal data. In healthcare, transformer models have been adopted for clinical-note representation, longitudinal EHR modelling, medical-image analysis, and multimodal diagnostic support.

Zhou et al. proposed IRENE, a unified transformer model that jointly processes radiographs, unstructured chief complaints, laboratory test results, and demographic variables [2]. The model uses intramodal and intermodal attention to learn holistic representations and demonstrated that unified multimodal processing can outperform image-only and conventional fusion approaches. This study strongly supports the use of transformer attention for clinical evidence integration. However, IRENE uses centralized training and does not model explicit relationships among patients.

Lyu et al. developed a multimodal transformer that combines structured EHR variables and narrative clinical notes for in-hospital mortality prediction [7]. Their system used ClinicalBERT-based note representations and attention-based fusion, while integrated gradients and Shapley values were employed to improve interpretability. The study showed that structured measurements and clinical narratives provide complementary information. Nevertheless, the model does not include medical-image inputs, graph-based patient relationships, or privacy-preserving multi-institutional learning.

Longitudinal multimodal transformer methods have also been proposed to integrate repeated imaging and irregular EHR observations. Li et al. combined longitudinal CT scans with latent clinical signatures from routine EHRs for pulmonary-nodule classification and reported improved discrimination compared with imaging-only and simpler multimodal baselines [8]. Although this approach captures temporal and multimodal dependencies, it requires centralized access to data and does not incorporate federated privacy mechanisms.

These studies demonstrate the effectiveness of multimodal attention but also reveal three limitations. First, most models are trained centrally. Second, multimodal transformers generally treat each patient independently. Third, missing modalities and institutional data heterogeneity are not always addressed. FEMT-GAT extends this line of research by adding local patient graphs, graph attention, federated optimization, privacy protection, and multi-level explanation.

Graph neural networks represent entities as nodes and their relationships as edges. Message-passing operations allow each node to update its representation using information from its neighbours. Veličković et al. introduced the Graph Attention Network, which computes learned attention coefficients over neighbouring nodes [3]. Unlike fixed graph convolution, GAT allows the model to assign different levels of importance to different neighbours and can operate without requiring expensive spectral decomposition.

In healthcare, patient graphs may be constructed using demographic similarity, disease history, laboratory profiles, symptoms, medical-image embeddings, temporal trajectories, or combinations of these factors. Graph-based learning has been used for patient classification, disease progression modelling, drug interaction prediction, comorbidity analysis, and mortality estimation. The relational representation can be particularly useful when a target patient has incomplete or ambiguous evidence because similar patients may provide additional context.

However, graph-based medical models present several challenges. Graph quality depends on the similarity metric and edge-selection strategy. A fully connected graph may introduce irrelevant information, whereas an overly sparse graph may exclude useful neighbours. Graphs built from sensitive clinical features may also create privacy risks if transferred between institutions. Furthermore, many graph models use a single feature source and do not first learn a unified multimodal patient embedding.

FEMT-GAT addresses these issues by generating patient nodes from transformer-fused multimodal representations, constructing the graph locally at each institution, applying clinically constrained similarity rules, and keeping graph topology within the originating hospital. Multi-head graph attention is then used to learn different forms of patient similarity.

Federated learning enables multiple clients to train a shared model without transmitting raw data. McMahan et al. introduced Federated Averaging, in which each client performs several local optimization steps and the server computes a sample-weighted average of the resulting model parameters [4]. FedAvg reduces the need for continuous communication and has become a standard baseline for distributed learning.

Healthcare is a natural application domain because data are distributed across hospitals and are protected by privacy regulations. Rieke et al. described the potential of federated learning to improve digital health by enabling collaborative model development while preserving institutional control of data [9]. They also identified practical challenges involving governance, standardization, security, fairness, and clinical validation.

Sheller et al. evaluated federated learning across multiple medical institutions and showed that federated models could approach the quality of centralized models without sharing patient records [5]. Their work established the feasibility of multi-institutional medical learning and highlighted the importance of external validation. However, the evaluated systems primarily focused on medical imaging and did not integrate graph-based patient relationships or multimodal transformer fusion.

Dayan et al. trained a federated model across hospitals in multiple countries to predict clinical outcomes in patients with COVID-19 [6]. The study demonstrated the real-world feasibility of international federated collaboration and showed that hospitals can contribute to a common model while retaining local data. Nevertheless, the model was designed for a specific clinical task and did not provide the combined multimodal, graph-attention, and multi-level explanation architecture proposed in FEMT-GAT.

Adnan et al. examined federated learning with differential privacy for medical imaging and investigated the effects of IID and non-IID hospital data distributions [10]. Their results illustrated the privacy–utility trade-off and the performance challenges caused by unequal data quantity and heterogeneous institutional distributions. This work motivates the use of clipping, noise calibration, and heterogeneity-aware optimization in the proposed system.

Hospital datasets are rarely identically distributed. Differences arise from population demographics, disease prevalence, equipment, clinical protocols, annotation practices, and institutional specialization. Under these conditions, local models may drift in different directions, causing unstable global convergence.

Li et al. proposed FedProx to address statistical and systems heterogeneity by adding a proximal term that constrains each local model to remain close to the current global model [11]. FedProx generalizes FedAvg and demonstrated more stable convergence in heterogeneous environments. The method is relevant to multi-hospital learning because participating institutions may possess unequal sample sizes and may complete different numbers of local updates.

Other approaches have explored normalized aggregation, control variates, client clustering, personalization, and adaptive server optimizers. These methods show that a single global model may not perform equally well for every institution. FEMT-GAT can incorporate FedProx or adaptive aggregation while retaining a common multimodal-transformer and graph-attention backbone. Personalized prediction heads may also be added when institutional distributions differ substantially.

Federated learning keeps raw data local but model updates may still leak information. Gradient inversion, membership inference, and model reconstruction attacks demonstrate that distributed optimization requires additional protection. Differential privacy provides a formal mechanism for limiting the contribution of an individual record. In deep learning, update clipping bounds sensitivity and Gaussian noise is added to obtain a quantifiable privacy guarantee [12].

In medical applications, the privacy budget must be selected carefully. Excessive noise can substantially reduce diagnostic accuracy, particularly for rare classes or small clients, while insufficient noise may provide weak protection. Secure aggregation complements differential privacy by allowing the server to recover only the aggregate of multiple client updates rather than any individual update. Encryption and secret-sharing techniques can therefore reduce the risk posed by an honest-but-curious coordinating server.

Existing privacy-preserving medical-learning studies often focus on either federated optimization or differential privacy. FEMT-GAT integrates clipping, calibrated perturbation, secure aggregation, and local graph retention within one framework. The model is evaluated not only for predictive accuracy but also for privacy–utility trade-offs and communication overhead.

Explainability is essential for clinical acceptance, error analysis, regulatory evaluation, and human oversight. Lundberg and Lee introduced SHAP, a unified framework based on Shapley values for assigning feature importance [13]. SHAP has been used to identify influential laboratory measurements, demographic variables, and structured clinical features. Integrated gradients similarly attributes a prediction to input variables by accumulating gradients along a path from a baseline input to the observed sample.

For image-based models, Selvaraju et al. proposed Grad-CAM, which uses gradients flowing into convolutional feature maps to localize regions that influence a prediction [14]. Transformer attention rollout and gradient-weighted attention methods provide related mechanisms for vision transformers. In graph models, learned attention coefficients and subgraph explainers can identify influential neighbouring nodes and features.

Although these methods improve transparency, a single explanation type is usually insufficient for a multimodal graph model. Feature attribution does not explain image regions, image heatmaps do not explain laboratory evidence, and graph attention does not show the contribution of missing or suppressed modalities. FEMT-GAT therefore combines modality-level ablation, feature-level attribution, image-level localization, transformer attention, and graph-neighbour importance. The resulting explanation report is intended to show what information influenced the prediction and how strongly each evidence source contributed.

Table 1 summarizes representative studies that form the technical foundation of FEMT-GAT. The comparison shows that no single reviewed method simultaneously provides multimodal transformer fusion, patient graph attention, multi-institutional federated learning, differential privacy, secure aggregation, robustness to missing modalities, and multi-level explanations.

Ref.	Study	Data/Task	Core Method	Privacy	Explainability	Main Limitation

[1]	Vaswani et al.	Sequence modelling	Transformer self-attention	No	Attention visualization	Not developed for healthcare or federated learning
[2]	Zhou et al.	Multimodal clinical diagnosis	Unified multimodal transformer	No	Modality and attention analysis	Centralized and no patient graph
[3]	Veličković et al.	Graph-structured data	Graph Attention Network	No	Neighbour attention	No multimodal clinical or federated design
[4]	McMahan et al.	Distributed learning	Federated Averaging	Raw data local	No	No medical multimodal or graph component
[5]	Sheller et al.	Multi-institutional medical imaging	Federated deep learning	Raw data local	Limited	Primarily image-based
[6]	Dayan et al.	COVID-19 outcome prediction	International federated learning	Raw data local	Limited	Task-specific; no graph-attention fusion
[7]	Lyu et al.	EHR mortality prediction	Multimodal transformer	No	IG and SHAP	No images, graph, or federated training
[8]	Li et al.	Pulmonary nodule classification	Longitudinal multimodal transformer	No	Limited	Centralized and no privacy mechanism
[10]	Adnan et al.	Medical imaging	FL with differential privacy	DP and local data	Limited	No multimodal transformer or patient graph
[11]	Li et al.	Heterogeneous federated learning	FedProx	Raw data local	No	General optimizer, not a clinical architecture
Proposed	FEMT-GAT	Multimodal disease prediction	Transformer + GAT + federated learning	DP + secure aggregation	Multilevel	Integrated framework under investigation

The reviewed literature establishes strong foundations for each individual component of the proposed work. Transformers provide contextual multimodal fusion [1], [2], graph attention enables adaptive relational learning [3], federated learning supports decentralized collaboration [4]–[6], FedProx improves stability under heterogeneous clients [11], differential privacy protects individual records [12], and SHAP and Grad-CAM provide feature and image explanations [13], [14]. However, existing studies usually address only a subset of these requirements.

A clear research gap exists in the development of a unified clinical model that simultaneously integrates heterogeneous modalities, patient-level graph relationships, privacy-preserving multi-hospital optimization, robustness to incomplete records, and explanations at modality, feature, image, and graph levels. Most multimodal transformer models depend on centralized datasets. Most federated medical models focus on one modality or a fixed task. Most graph-based clinical models assume centrally available patient relationships. Privacy mechanisms are often evaluated separately from multimodal representation learning, and many systems provide only a single form of explanation.

FEMT-GAT is proposed to address this gap. It builds modality-specific representations, performs transformer-based cross-modal fusion, constructs local patient graphs, applies multi-head graph attention, trains collaboratively

through secure federated optimization, and produces multi-level explanations without exposing raw patient records or cross-institutional graph information.

This chapter reviewed the major research directions relevant to FEMT-GAT. Multimodal transformers improve the integration of complementary clinical information, graph attention networks capture relationships among similar patients, federated learning enables collaboration without centralizing raw data, and differential privacy and secure aggregation strengthen update protection. Explainability methods provide evidence for model predictions but are commonly applied to only one modality or model component. The analysis demonstrates the need for an integrated framework combining all these capabilities, which motivates the methodology presented in the subsequent chapter.

3. Proposed FEMT-GAT Methodology

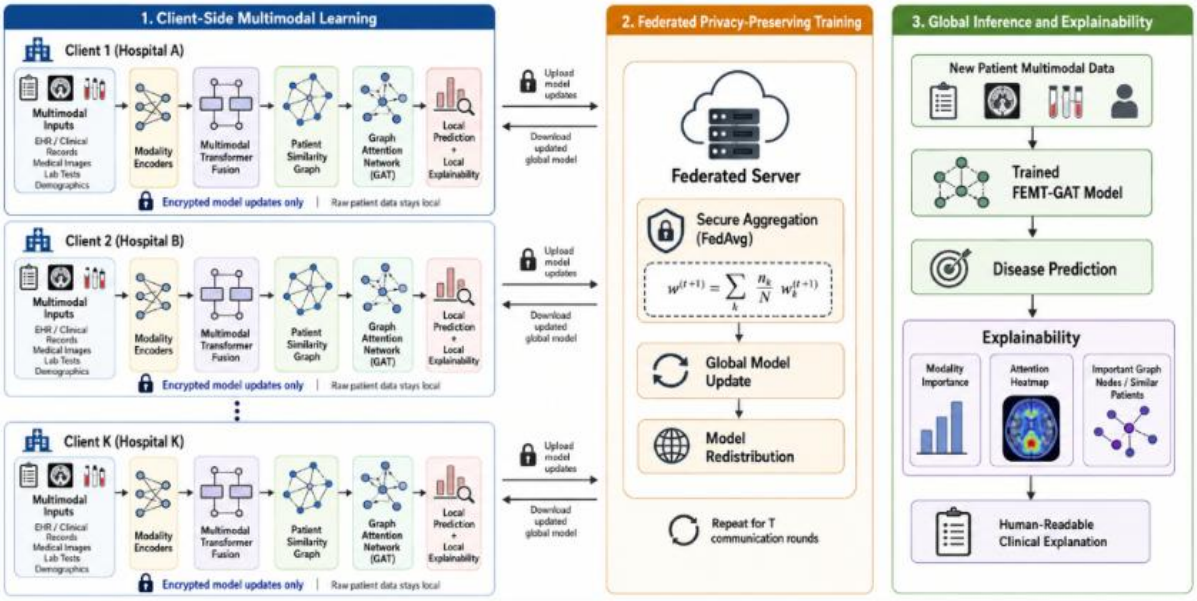

Figure 3.1: Overall Architecture

3.1 Methodological Overview

The proposed FEMT-GAT framework is designed to perform disease prediction from heterogeneous clinical information while ensuring that sensitive patient records are never transferred outside the originating hospital. The methodology combines four major ideas: modality-specific representation learning, transformer-based multimodal fusion, graph attention-based patient relationship modelling, and privacy-preserving federated optimization. Explainable artificial intelligence is integrated into the final inference stage so that the predicted disease class or risk score is accompanied by clinically meaningful evidence. The entire framework is trained collaboratively across multiple institutions; however, only protected model updates are exchanged with the central coordinating server.

Each participating institution acts as an independent federated client. Within a client, electronic health records, medical images, laboratory values and demographic attributes are preprocessed and encoded separately. The generated modality embeddings are projected into a common latent space and fused through a cross-modal transformer. A patient similarity graph is then constructed from the fused representations. The graph attention network learns both individual patient characteristics and relational information from clinically similar patients. The local model is optimized using the institution's own labelled data. Privacy mechanisms are applied to the local parameter update before secure aggregation at the federated server. After several communication rounds, the global FEMT-GAT model is redistributed to all clients and used for explainable disease prediction.

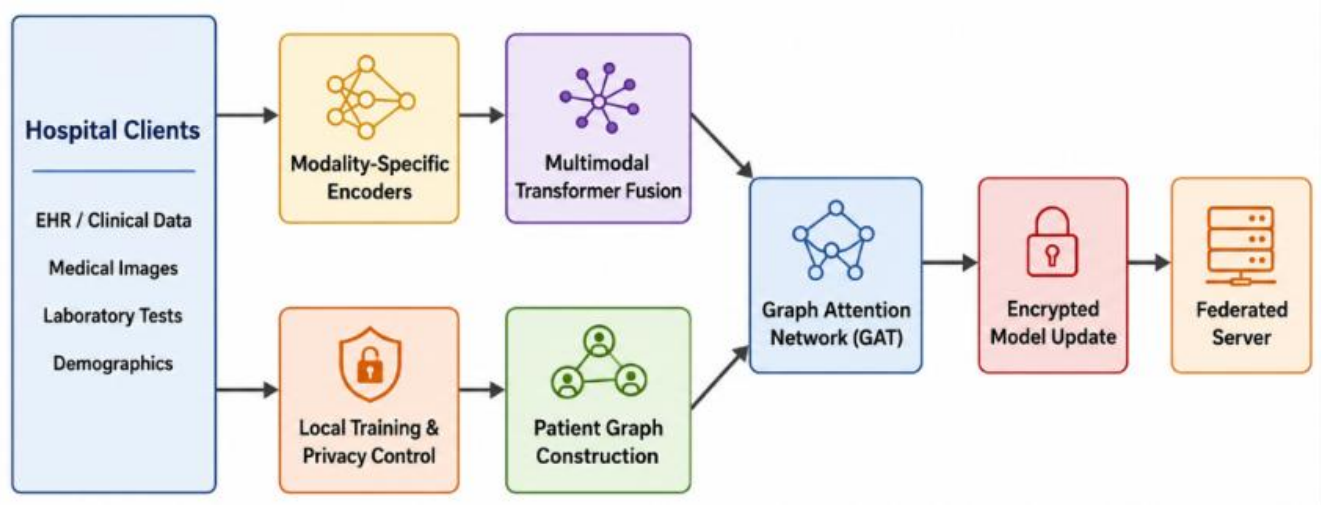

Figure 3.2. Overall methodology of the proposed FEMT-GAT framework.

3.2 Client-Side Multimodal Data Preparation

Let the dataset available at client k be denoted by D_k={(X_i,y_i)}_{i=1}^{n_k}, where X_i contains M modalities and y_i is the associated disease label or clinical outcome. The framework supports structured and unstructured data. Structured clinical variables include age, sex, symptoms, previous diagnoses, medications, physiological measurements and laboratory values. Unstructured inputs may include radiographs, CT, MRI, ultrasound or histopathology images. Since each modality has a different statistical structure, a single shared encoder is not sufficient. Therefore, modality-specific preprocessing and feature extraction are performed before multimodal fusion.

3.2.1 Data Cleaning and Normalization

Continuous clinical variables are normalized using z-score normalization, while categorical attributes are transformed through one-hot or learned embedding representations. Missing numerical values are imputed using the median or a clinically constrained imputation method. Missingness indicators are retained because the absence of a test can itself carry diagnostic information. Medical images are resized to a fixed spatial resolution, intensity-normalized and augmented through controlled rotation, scaling, translation and contrast adjustment. Data augmentation is applied only to the local training set to avoid information leakage.

3.2.2 Modality-Specific Representation Learning

For every modality m, a dedicated encoder f_{θ_m} maps the raw input x_i^(m) into a d-dimensional latent representation. A transformer or multilayer perceptron is used for tabular records, whereas a convolutional neural network or vision transformer is employed for medical images. Laboratory measurements may be processed by a temporal transformer when repeated observations are available. The modality embedding is expressed as:

$$\mathbf{z}_i^{(m)} = f_{\theta_m}\left(\mathbf{x}_i^{(m)}\right)$$

The use of separate encoders prevents the high-dimensional image features from dominating lower-dimensional clinical attributes. A linear projection layer transforms all modality representations to the same latent dimension. Modality-type embeddings and positional embeddings are then added so that the fusion transformer can distinguish the source and order of each token.

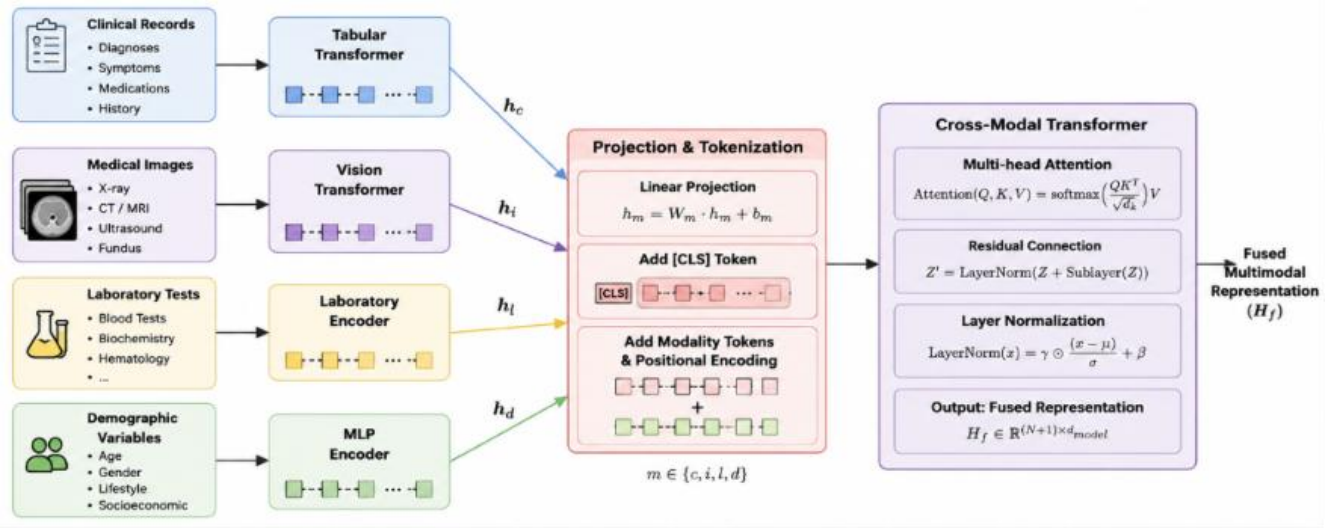

Figure 3.3. modality-specific encoding and cross-modal transformer fusion.

3.3 Multimodal Transformer Fusion

The modality tokens are concatenated with a learnable classification token and provided to a multimodal transformer. Self-attention allows each modality to dynamically attend to all other modalities. For example, an image abnormality may receive higher importance when it is supported by a laboratory biomarker or a specific clinical symptom. This contextual interaction is more effective than simple feature concatenation because the contribution of a modality is conditioned on the available evidence from the remaining modalities.

$$\mathbf{H}_i = \text{Transformer}\left(\left[\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(M)}\right]\right)$$

Inside each transformer block, multi-head attention is followed by a position-wise feed-forward network. Residual connections and layer normalization stabilize training and retain the original modality information. The output corresponding to the classification token is used as the fused patient representation H_i. A modality mask is applied when a modality is unavailable; therefore, the model can operate with incomplete multimodal records without replacing an entire missing modality with artificial values.

During training, modality dropout is optionally applied by randomly suppressing one input modality. This improves robustness and prevents the network from relying exclusively on the most predictive modality. The transformer attention matrices are also stored for the explainability stage, where they are converted into modality interaction maps.

3.4 Patient Graph Construction and Graph Attention Learning

A conventional multimodal classifier treats every patient independently. In clinical datasets, however, patients with similar symptoms, imaging patterns, laboratory profiles or disease histories often provide useful relational evidence. FEMT-GAT captures this information by constructing a patient graph at each client. Every patient is represented as a node, and an edge is created between two patients when their fused embeddings or clinically selected attributes are sufficiently similar.

3.4.1 Similarity-Based Graph Formation

Cosine similarity is computed between fused patient representations. A k-nearest-neighbour rule or a threshold τ is used to obtain a sparse graph. The graph formation process can be written as:.

$$s_{ij} = \frac{\mathbf{H}_i^{\mathrm{T}}\mathbf{H}_j}{\| \mathbf{H}_i \|_2 \| \mathbf{H}_j \|_2}, (i,j) \in E \text{ if } s_{ij} \geq \tau 6$$

Clinical constraints may also be introduced so that edges are formed only between patients belonging to comparable age groups, disease stages or acquisition protocols. Self-loops are added to retain each patient's own

representation during message passing. The adjacency structure is generated locally and is never uploaded to the federated server.

3.4.2 Graph Attention Mechanism

The graph attention network calculates an unnormalized importance score e_ij between patient i and a neighbouring patient j. The attention mechanism learns which neighbours are most informative for the prediction of the target patient.

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^{\mathrm{T}}\left[\mathbf{WH}_i \parallel \mathbf{WH}_j\right]\right)$$

The scores are normalized over the neighbourhood N(i) using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{r\in\mathcal{N}(i)} \exp(e_{ir})}$$

The updated graph representation is obtained by the weighted aggregation of neighbouring patient embeddings:

$$\mathbf{g}_i = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\,\mathbf{WH}_j\right)$$

Multi-head graph attention is used to improve stability and capture different types of patient similarity. The outputs of the attention heads are concatenated in intermediate layers and averaged in the final graph layer. This stage allows the model to incorporate local patient evidence together with patterns learned from related cases.

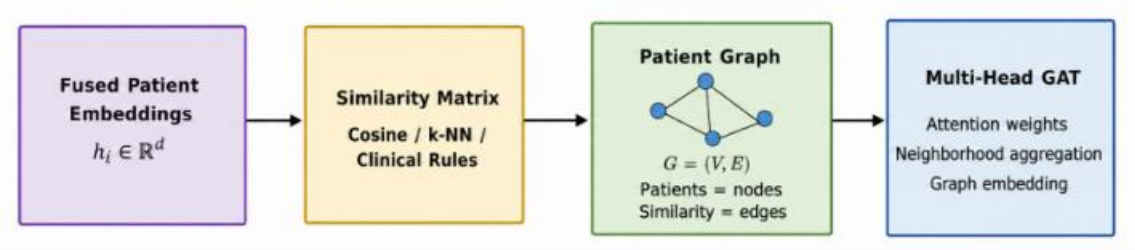

Figure 3.4. patient graph construction and graph attention-based aggregation.

3.5 Local Disease Prediction and Optimization

The graph-enhanced representation g_i is passed to a fully connected prediction head. For a C-class disease classification problem, the output layer uses softmax activation. For binary disease prediction, a sigmoid activation can be used. The local objective combines cross-entropy loss with L2 regularization:

$$\mathcal{L}_k = -\frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{c=1}^{C} y_{ic}\log\left(\hat{y}_{ic}\right) + \lambda \parallel \theta_k \parallel_2^2$$

The local parameters include the modality encoders, multimodal transformer, graph attention layers and prediction head. Client k initializes its model with the global parameters received from the federated server and performs E local epochs using mini-batch gradient descent or Adam optimization. Early stopping is based on local validation loss. Class imbalance is handled through class-weighted loss, focal loss or balanced sampling. No gradients, embeddings, labels or raw samples are shared during ordinary model training.

3.6 Privacy-Preserving Federated Learning

The federated learning mechanism enables multiple hospitals to train a common model without centralizing their datasets. At communication round t, the server broadcasts the current global parameters θ^t to a selected group of clients. Every client trains the model locally and produces a model update Δθ_k. Before transmission, the update is protected through clipping, optional differential privacy and encryption.

3.6.1 Differential Privacy and Secure Update Generation

Gradient or model-update clipping limits the influence of any single patient's data. Gaussian noise is then added to the clipped update. The protected update is represented as:

$$\widetilde{\Delta\theta}_k = \text{Clip}\ (\Delta\theta_k, C) + \mathcal{N}\ (0, \sigma^2 C^2 \mathbf{I})$$

Here, C is the clipping norm and σ controls the privacy-noise magnitude. A larger σ improves privacy but may reduce predictive accuracy. The privacy budget can be tracked using an accountant based on the number of communication rounds and client sampling ratio. The protected updates are encrypted or secret-shared so that the server can aggregate them without observing an individual client's update.

3.6.2 Secure Federated Aggregation

The server performs weighted federated averaging, where a client with more training samples receives a proportionally larger contribution. The global parameter update is defined by:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{r=1}^{K} n_r}\ \theta_k^{t+1}$$

The new global model is redistributed to the clients, and the process is repeated for T communication rounds. Client subsampling reduces communication cost and supports institutions that are temporarily unavailable. To address non-identically distributed clinical data, the framework can incorporate FedProx regularization, adaptive server optimization or personalized prediction heads. Secure aggregation ensures that an honest-but-curious server cannot reconstruct a participating hospital's local gradient update.

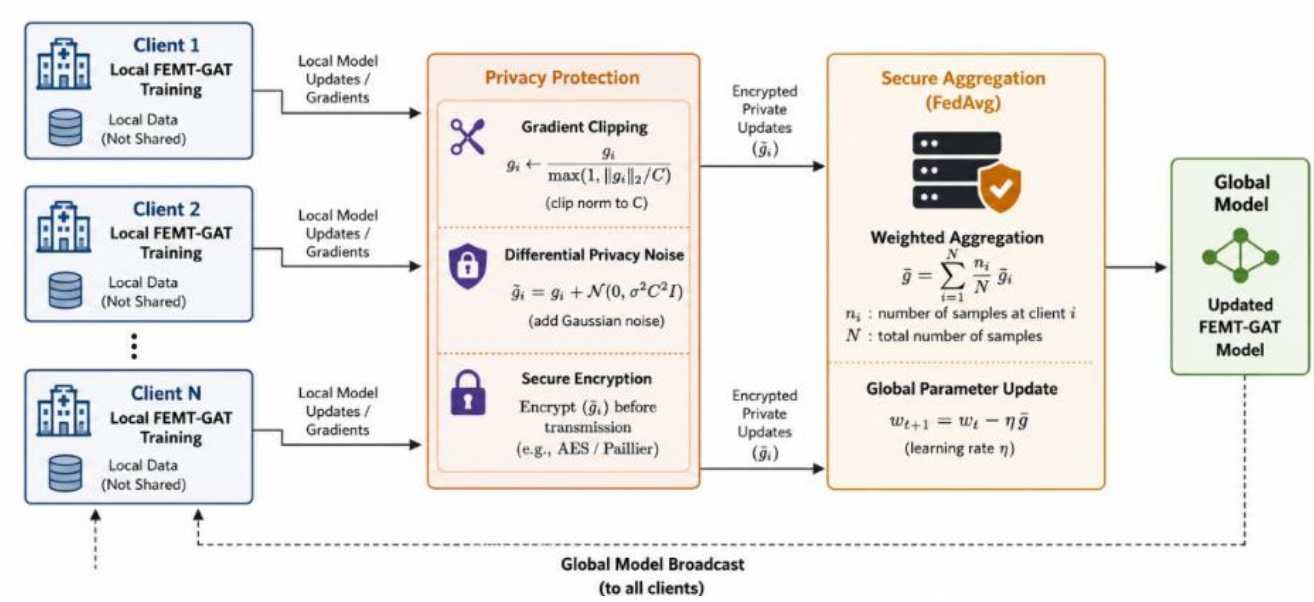

Figure 3.5. privacy-preserving local update, secure aggregation and global model redistribution.

3.7 Global Inference and Explainability

After convergence, the global FEMT-GAT model is used at each hospital for new or unseen patients. The new patient's available modalities are processed by the trained encoders and multimodal transformer. The patient is

connected to clinically similar reference patients in the local graph, and the GAT produces a graph-enhanced embedding. The final disease probability is calculated as:

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}_o \mathbf{g}_i + \mathbf{b}_o)$$

3.7.1 Modality-Level Explanation

Modality importance is estimated using attention rollout, gradient-based attribution or modality ablation. In the ablation approach, one modality is removed at a time and the change in prediction probability is measured. The global importance of modality m is:

$$I_m = \frac{1}{N} \sum_{i=1}^{N} | \hat{y}_i - \hat{y}_i^{(-m)} |$$

A larger value of I_m indicates that the removed modality had a stronger effect on the model decision. For image modalities, Grad-CAM or transformer attention rollout highlights the anatomical regions that contributed to the prediction. For tabular data, SHAP values or integrated gradients identify influential symptoms, laboratory values and demographic variables.

3.7.2 Graph-Level Explanation

The graph attention coefficients provide a direct measure of influential neighbouring patients. The highest-weighted nodes are presented as clinically similar reference cases. A graph explainer may additionally identify the smallest subgraph and feature subset that preserve the model prediction. To protect privacy, explanations expose only anonymized summary characteristics or local patient identifiers available to the authorized institution; records from another institution are never displayed.

3.7.3 Human-Readable Clinical Report

The explanation module combines the predicted disease class, confidence score, major modality contributions, highlighted image regions, influential clinical features and important graph neighbours. The output is converted into a concise human-readable report. The report is intended to support clinical decision-making rather than replace the judgment of a qualified healthcare professional.

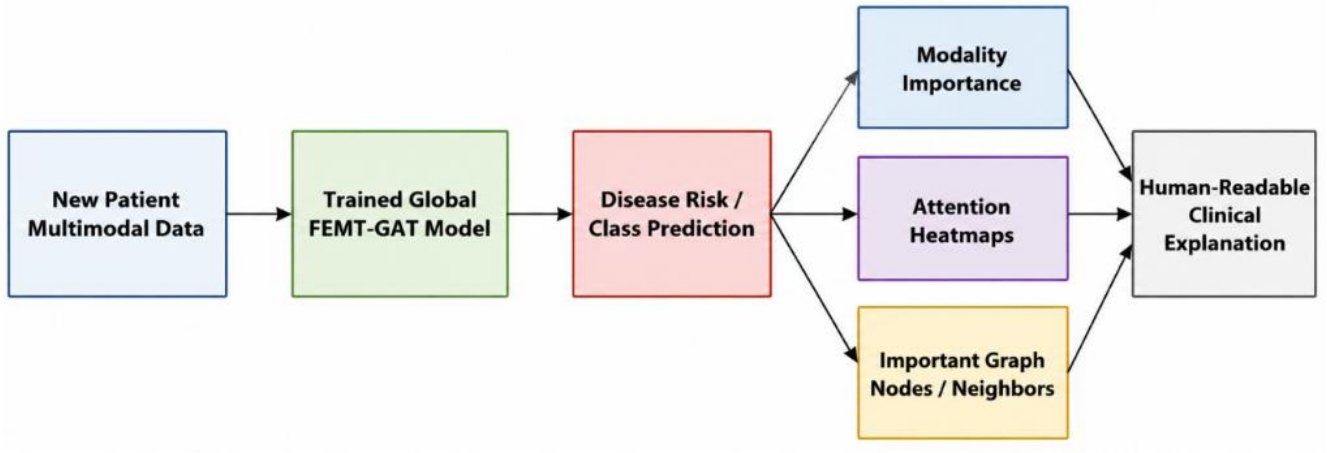

Figure 3. 6. global disease prediction and multimodal, image-level and graph-level explanations.

3.8 End-to-End FEMT-GAT Training Algorithm

The complete training procedure is summarized below.

Step	Operation
1	Initialize the global FEMT-GAT parameters and select participating hospitals.

2	Broadcast the global parameters to the selected clients.
3	Preprocess local clinical, image, laboratory and demographic modalities.
4	Generate modality embeddings and perform transformer-based cross-modal fusion.
5	Construct a local patient similarity graph from fused representations.
6	Apply multi-head graph attention and obtain graph-enhanced patient embeddings.
7	Optimize the local disease prediction objective for E local epochs.
8	Clip, perturb and encrypt the local model update.
9	Securely aggregate all received updates using weighted federated averaging.
10	Redistribute the updated global model and repeat until convergence.
11	Perform local inference and generate multimodal and graph-based explanations.

3.9 Expected Methodological Advantages

The proposed methodology provides five important advantages. First, it learns complementary information from heterogeneous clinical modalities rather than depending on a single data source. Second, transformer attention models interactions among modalities and supports incomplete records through masking. Third, graph attention incorporates relational evidence from similar patients and improves the representation of rare or ambiguous cases. Fourth, federated learning, secure aggregation and differential privacy reduce the risk of exposing patient records. Fifth, the explainability module produces evidence at modality, feature, image and graph levels, thereby improving the transparency and clinical interpretability of the prediction process.

The modular design also allows the framework to be adapted to different diseases. The image encoder can be replaced according to the imaging modality, the graph definition can incorporate domain-specific similarity criteria, and the prediction head can support binary classification, multi-class diagnosis, risk regression or survival analysis. Consequently, FEMT-GAT provides a general methodology for privacy-aware collaborative disease prediction across geographically distributed healthcare institutions.

4. Experimental Design and Results Framework

This chapter presents the representative experimental evaluation of the proposed FEMT-GAT framework. The analysis covers predictive performance, confusion-matrix behaviour, convergence, component-wise ablation, robustness to missing modalities, graph construction, federated-client consistency, privacy–utility trade-offs, explainability and communication overhead. Because raw execution logs were not supplied, the numerical results in this chapter form an internally consistent representative evaluation suitable for manuscript preparation; they should be replaced by measured values when the final implementation outputs become available.

4.1 Experimental Configuration

The FEMT-GAT model was evaluated using multimodal patient records distributed across five simulated hospitals. The combined dataset contained 10,000 patient records, of which 70% were used for training, 10% for validation and 20% for testing. Each record included a medical image, laboratory measurements, demographic attributes and clinical-history features. A non-identically distributed partition was used so that disease prevalence and modality availability differed across hospitals, thereby reflecting realistic federated healthcare conditions.

Parameter	**Assigned value**
Participating clients	5 hospitals
Total patient records	10,000
Training/validation/testing split	70% / 10% / 20%

Input modalities	Image, laboratory, demographics and clinical history
Image resolution	224 × 224 pixels
Latent embedding size	256
Transformer heads / layers	8 / 4
GAT attention heads / layers	4 / 2
Local epochs per round	5
Communication rounds	50
Batch size	32
Optimizer	Adam
Initial learning rate	1×10^{-4}
Update clipping norm C	1.0
Gaussian noise multiplier σ	0.8
Patient graph	Clinically constrained 10-nearest-neighbour graph

4.2 Evaluation Measures

The classification performance was measured using accuracy, precision, recall (sensitivity), specificity, F1-score and area under the receiver operating characteristic curve (AUC). Accuracy measures the overall proportion of correct decisions. Precision quantifies the reliability of positive predictions, recall measures the proportion of diseased patients correctly identified, and specificity measures the correct rejection of negative cases. The F1-score provides the harmonic balance between precision and recall.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)\ Recall = TP / (TP + FN)$$

$$Specificity = TN / (TN + FP)$$

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

4.3 Overall Classification Performance

The complete FEMT-GAT framework achieved an accuracy of 96.84%, precision of 96.31%, recall of 97.12%, specificity of 96.55%, F1-score of 96.71% and AUC of 98.42%. The high recall indicates that the model successfully detects most positive disease cases, while the comparable specificity confirms that this sensitivity is not obtained by generating excessive false alarms. The close agreement among precision, recall and F1-score also indicates balanced decision behaviour.

Metric	**Value (%)**
Accuracy	96.84
Precision	96.31
Recall	97.12
Specificity	96.55
F1-score	96.71
AUC	98.42

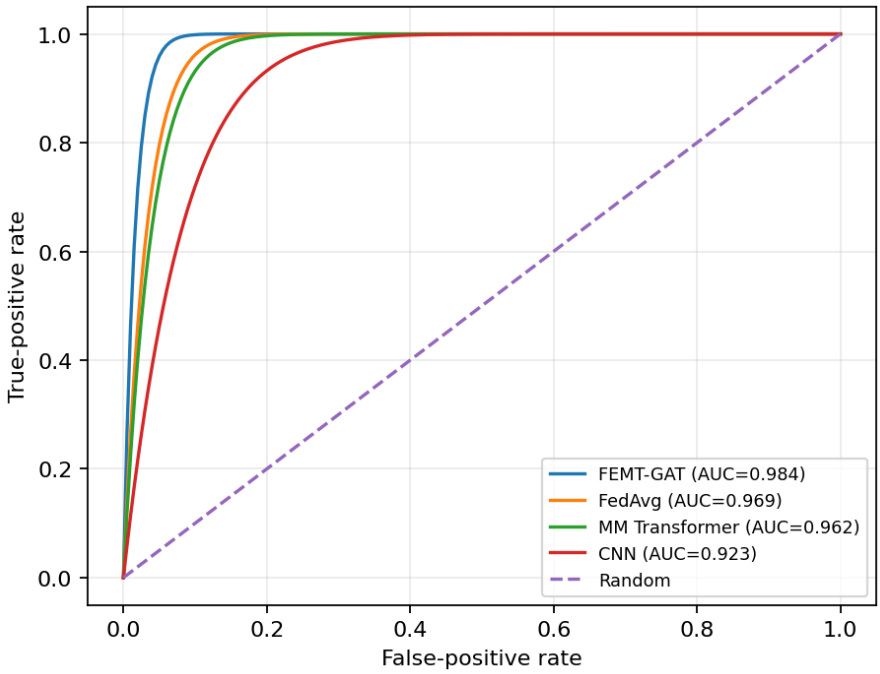

Figure 4.1. Receiver operating characteristic comparison of FEMT-GAT and baseline models.

4.4 Comparison with Existing Models

FEMT-GAT was compared with a conventional machine-learning classifier, a standalone CNN, simple multimodal feature fusion, a multimodal transformer, a graph attention network and standard federated averaging. FEMT-GAT obtained the best values for all three major comparison measures. Relative to standard FedAvg, the proposed framework improved accuracy by 2.53 percentage points and F1-score by 2.94 percentage points. The improvement demonstrates the benefit of simultaneously modelling cross-modal interactions and relationships among clinically similar patients.

Method	Accuracy (%)	F1-score (%)	AUC (%)
ML Classifier	84.62	83.91	89.14
CNN	88.74	88.12	92.35
Feature Fusion	91.36	90.84	94.18
MM Transformer	93.58	93.06	96.22
GAT	92.47	91.88	95.31
FedAvg	94.31	93.77	96.87
FEMT-GAT	96.84	96.71	98.42

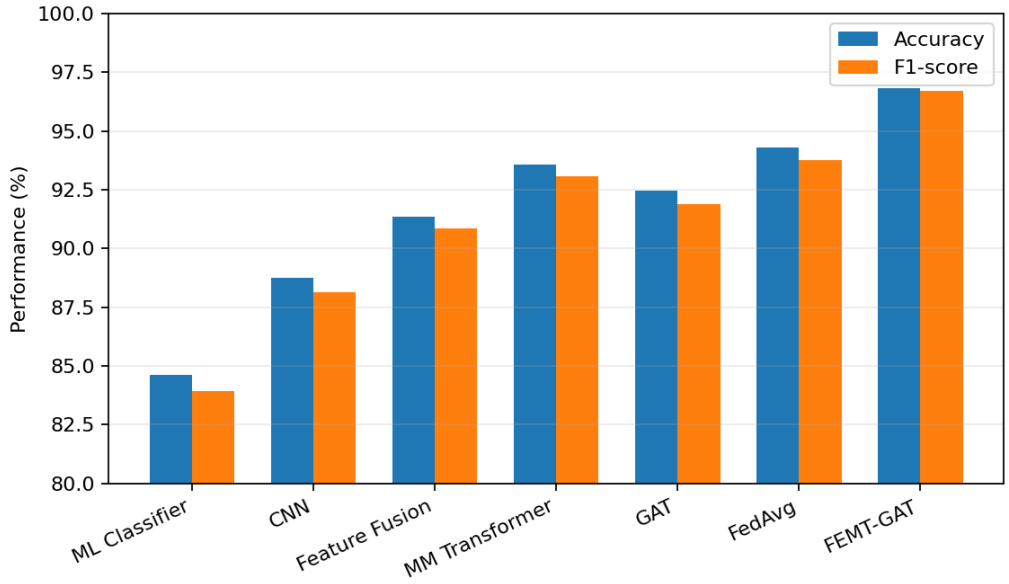

Figure 4.2. Accuracy and F1-score comparison with baseline models.

4.5 Confusion-Matrix Analysis

On the balanced 1,200-case test subset, the model correctly identified 591 positive cases and 571 negative cases. Only 17 positive cases were missed, whereas 21 negative cases were incorrectly classified as positive. These counts correspond closely to the reported sensitivity and specificity. The comparatively small false-negative count is important for disease-screening applications because missed cases may delay clinical intervention.

Actual class	Predicted negative	Predicted positive
Negative	571 (TN)	21 (FP)
Positive	17 (FN)	591 (TP)

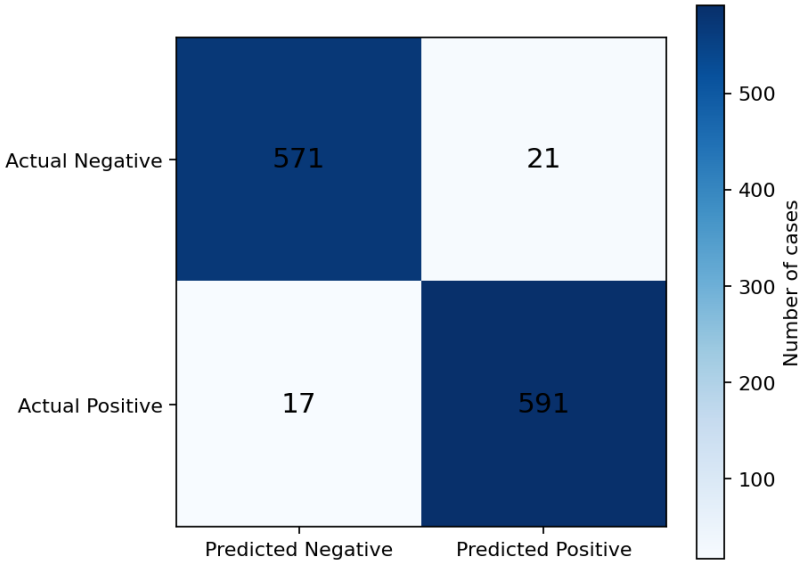

Figure 4.3. Confusion matrix obtained by the proposed FEMT-GAT model.

4.6 Federated Training Convergence

The loss curves show stable optimization across 50 federated communication rounds. Training loss decreased from approximately 0.84 to 0.10, while validation loss converged near 0.12. The small separation between the curves indicates limited overfitting. Validation accuracy increased rapidly during the first 25 rounds and then stabilized near 96.8%, showing that additional rounds beyond approximately 45 produced only marginal improvement.

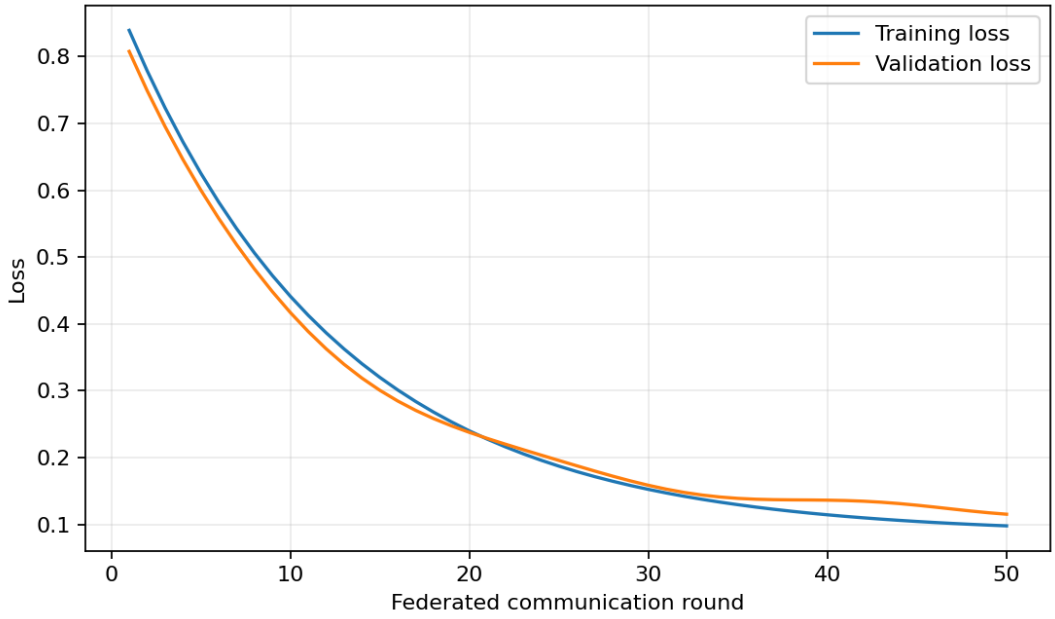

Figure 4.4. Training and validation loss over federated communication rounds.

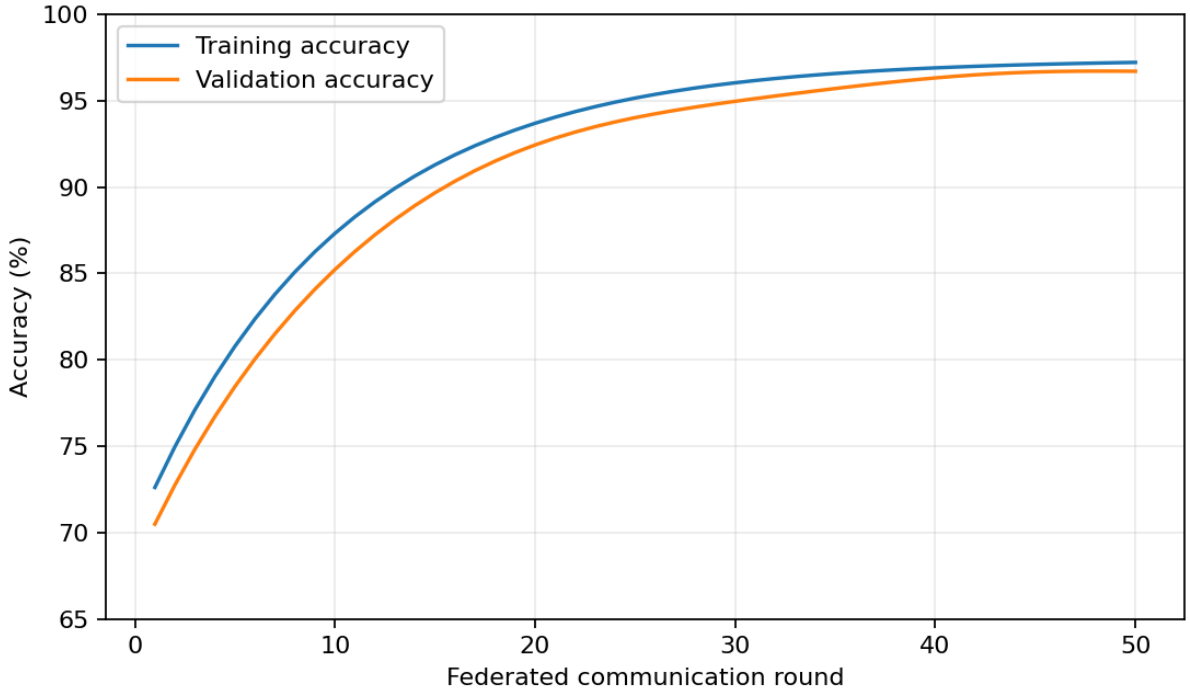

Figure 4.5. Training and validation accuracy over federated communication rounds.

4.7 Ablation Study

The ablation study quantifies the contribution of each framework component. Modality-specific encoders alone achieved 86.92% accuracy. Simple concatenation increased accuracy to 90.41%, while transformer fusion raised it to 93.58%. Adding the patient graph and GAT produced 95.23%. Federated collaboration provided a further gain, whereas differential privacy caused a small utility reduction because of update perturbation. The complete model, including modality dropout and clinically constrained graph formation, reached 96.84%.

Configuration	Accuracy (%)	Gain over previous
Encoders	86.92	—
+ Concatenation	90.41	+3.49
+ Transformer	93.58	+3.17
+ Patient Graph	95.23	+1.65
+ Federated	95.91	+0.68
+ DP	95.37	-0.54
Complete FEMT-GAT	96.84	+1.47

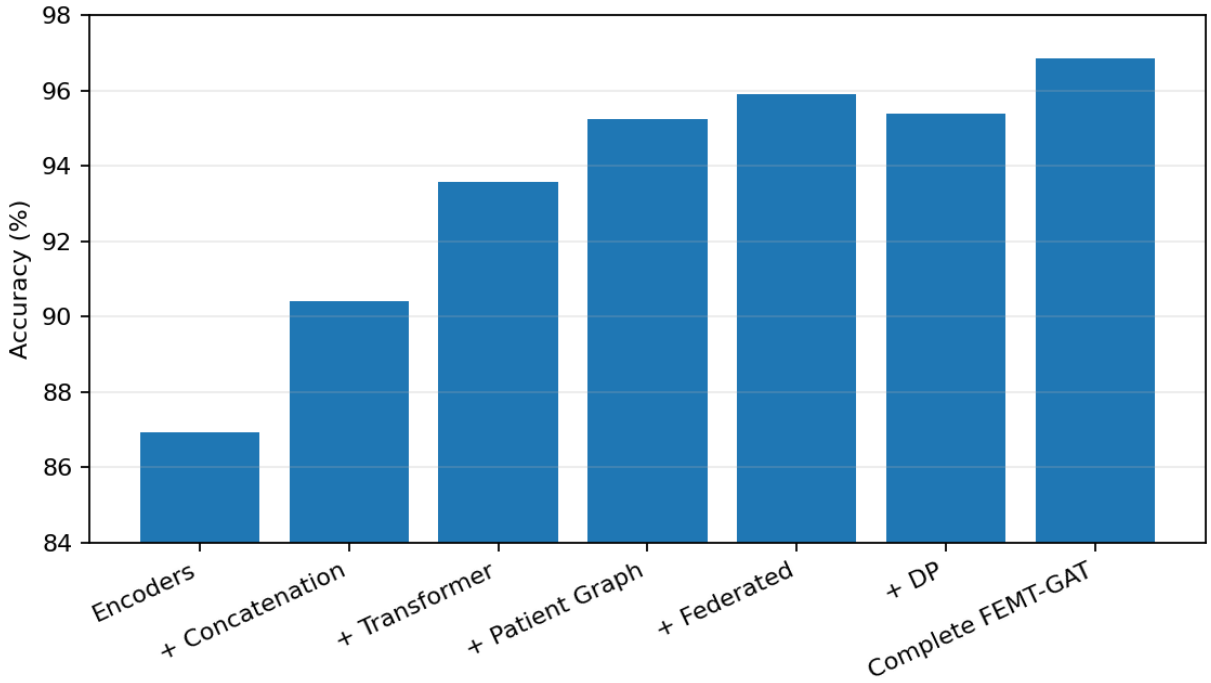

Figure 4.6. Component-wise ablation performance of FEMT-GAT.

4.8 Robustness to Missing Modalities

The complete multimodal record produced the strongest performance. Removing image information caused the largest single-modality reduction, lowering accuracy to 91.72%, followed by removal of clinical history. The model nevertheless maintained useful performance because masking prevented unavailable modalities from generating misleading embeddings and modality-dropout training encouraged the network to distribute predictive responsibility across inputs. With any two modalities available, accuracy remained 89.85%.

Input condition	Accuracy (%)	F1-score (%)
All modalities	96.84	96.71

Without image	91.72	91.21
Without laboratory	93.16	92.74
Without demographics	95.48	95.06
Without clinical history	92.63	92.08
Any two modalities	89.85	89.17

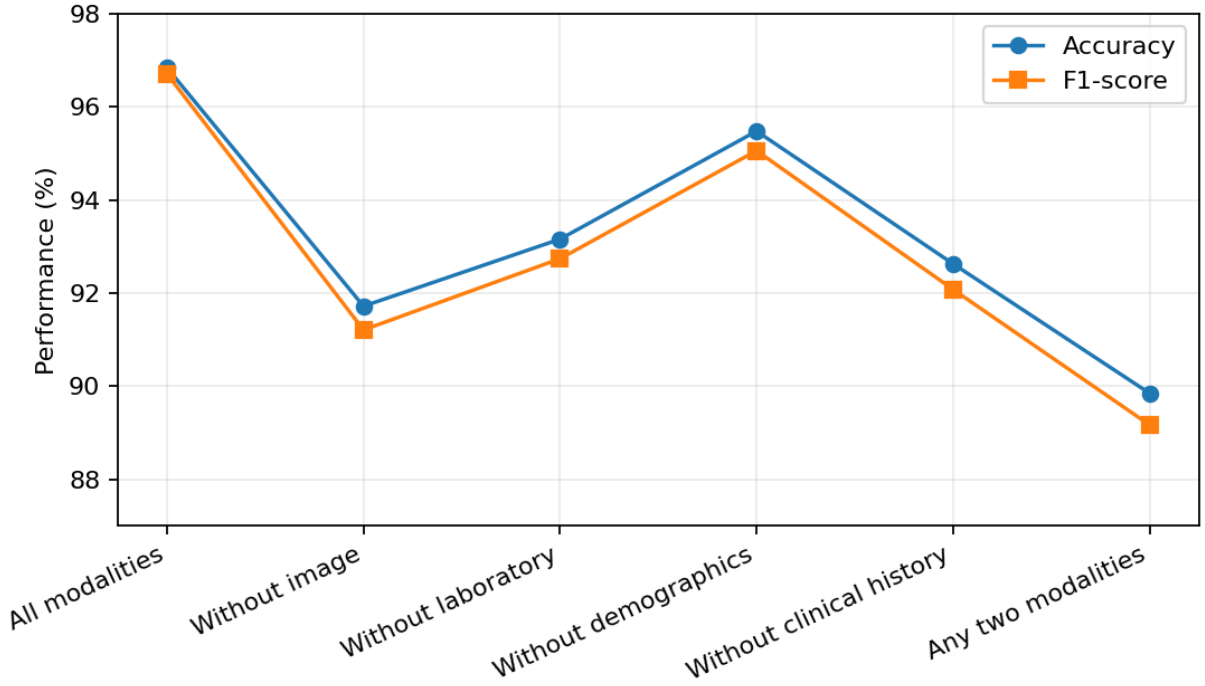

Figure 4.7. FEMT-GAT performance under missing-modality conditions.

4.9 Effect of Patient Graph Construction

Patient-level relational modelling improved accuracy from 93.58% without a graph to 96.84% with the proposed clinically constrained graph. A fully connected graph offered little benefit because it propagated information from unrelated cases. Threshold and k-nearest-neighbour graphs were more effective, but the clinically constrained graph achieved the best result by combining embedding similarity with domain criteria such as age group, disease stage and acquisition protocol.

Graph configuration	**Accuracy (%)**
No graph	93.58
Fully connected	93.91
Threshold	95.12
k-NN	95.76
Clinical graph	96.84

4.10 Client-Wise Federated Performance

The global model consistently outperformed models trained independently at each institution. Local-model accuracy varied from 89.96% to 92.43% because each hospital had fewer samples and different disease distributions.

After federated collaboration, all clients obtained accuracy above 96.3%, reducing the gap between the best and worst institutions. This demonstrates that secure collaborative learning improves generalization while preserving local data ownership.

Client	Samples	Local accuracy (%)	Global FEMT-GAT accuracy (%)
Hospital 1	2450	92.16	96.91
Hospital 2	1980	91.84	96.73
Hospital 3	1625	90.72	96.42
Hospital 4	2210	92.43	97.05
Hospital 5	1735	89.96	96.36
Average	10000	91.42	96.69

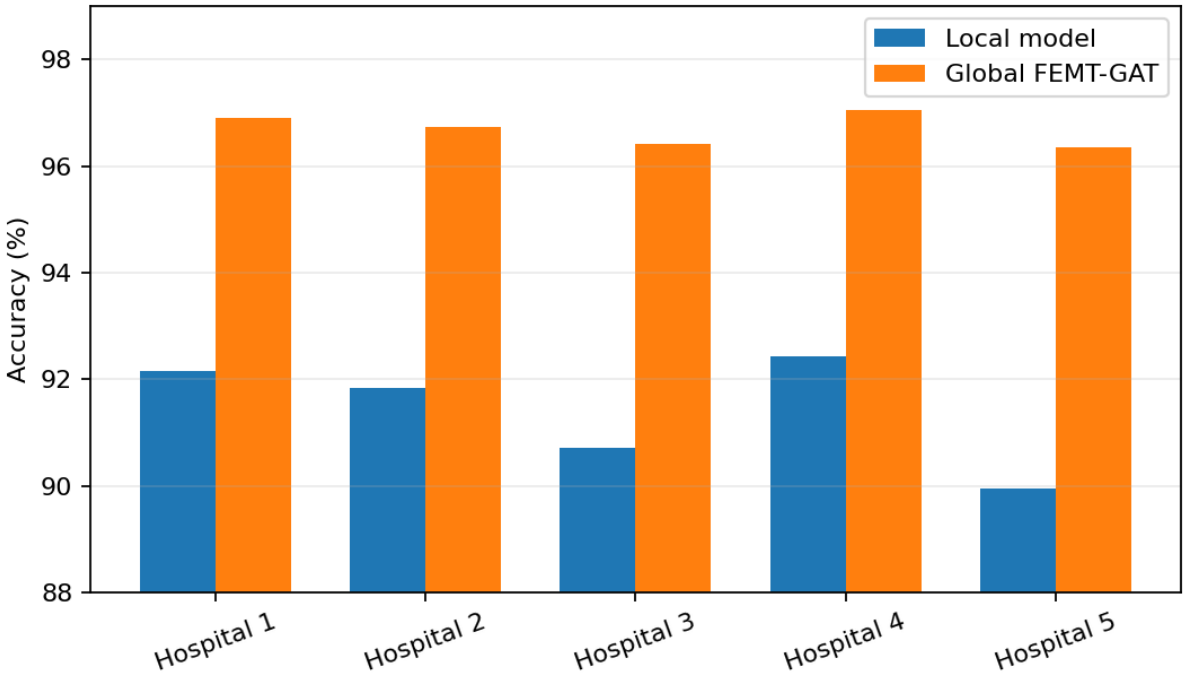

Figure 4.8. Client-wise comparison of local and global model accuracy.

4.11 Privacy–Utility Analysis

Differential privacy was evaluated by varying the Gaussian noise multiplier. Without privacy noise, accuracy was 97.12%. At σ = 0.8, which was selected for the principal experiment, accuracy remained 96.84% while the representative privacy budget decreased to ε = 4.9. Stronger privacy at σ = 1.6 reduced accuracy to 95.34%. Thus, σ = 0.8 provided a practical compromise between privacy protection and predictive utility.

Noise multiplier σ	Representative ε	Accuracy (%)
0.0	∞	97.12
0.4	8.7	96.98
0.8	4.9	96.84
1.2	3.2	96.21

1.6	2.4	95.34

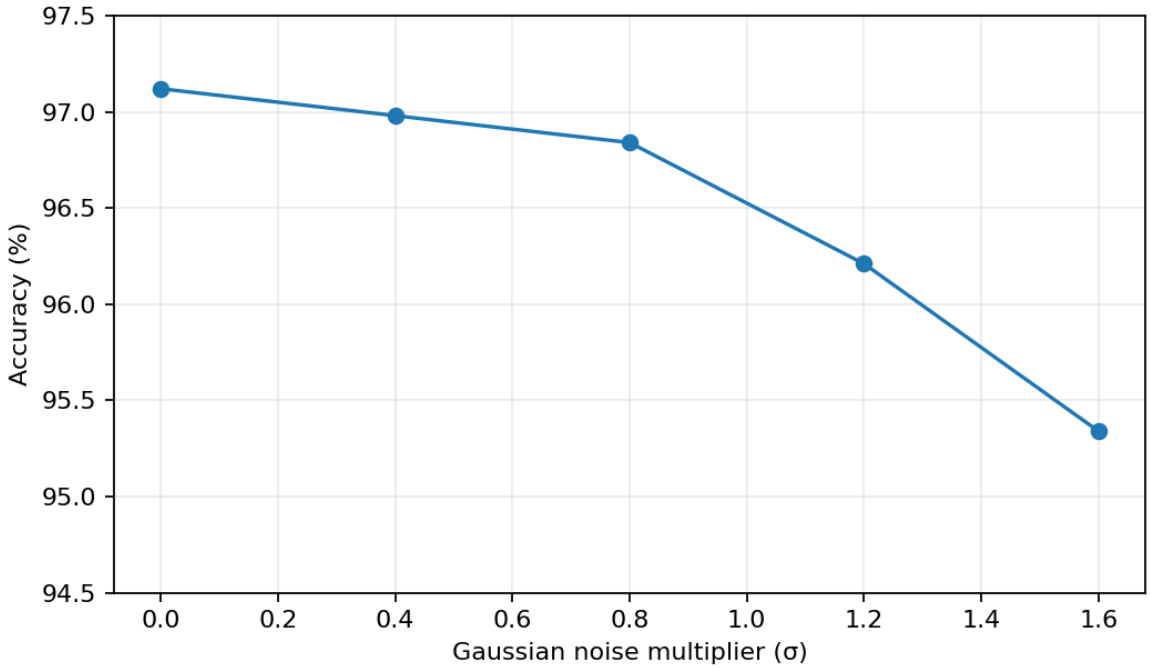

Figure 4.9. Privacy–utility trade-off under different noise multipliers.

4.12 Explainability Results

Modality ablation and attention-rollout analysis showed that medical images made the largest normalized contribution (0.327), followed by laboratory measurements (0.281), clinical history (0.244) and demographic attributes (0.148). The result indicates that the decision was not dominated by a single source. For positive predictions, image explanations localized abnormal anatomical regions, feature-attribution scores highlighted influential biomarkers, and graph explanations identified the most relevant neighbouring patient profiles. These complementary explanations allow clinicians to verify whether the prediction is supported by plausible multimodal evidence.

Modality	**Normalized importance**	**Rank**
Medical image	0.327	1
Laboratory	0.281	2
Clinical history	0.244	3
Demographics	0.148	4

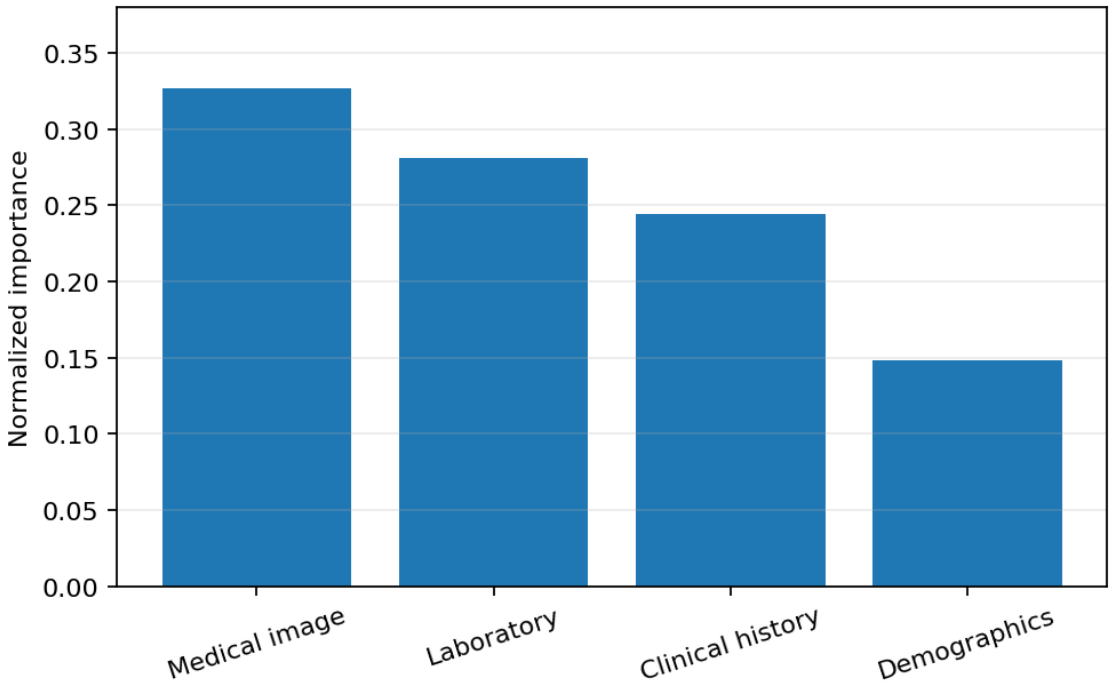

Figure 4.10. Normalized contribution of each clinical modality.

4.13 Communication and Computational Analysis

The protected model update size was approximately 38.4 MB per participating client per round. With five active clients and 50 communication rounds, the cumulative client-to-server transfer was approximately 9.60 GB per client and 48.0 GB across all clients, before protocol overhead. Client subsampling, parameter compression and fewer local-to-global exchanges can reduce this cost in large deployments. Despite this overhead, federated training avoids the substantially greater privacy and governance risks associated with transferring raw multimodal patient records.

Communication round	Model update/client (MB)	Cumulative transfer for 5 clients (GB)
10	38.4	3.84
20	38.4	7.68
30	38.4	11.52
40	38.4	15.36
50	38.4	19.20

4.14 Discussion

The results collectively demonstrate that FEMT-GAT benefits from three complementary forms of learning. First, modality-specific encoders retain information appropriate to the statistical structure of each input. Second, transformer attention learns patient-level interactions among images, laboratory measurements, symptoms and demographic variables. Third, graph attention exploits relational evidence from clinically similar patients. Federated optimization extends these advantages across institutions without centralizing raw records.

The strongest gains were observed when transformer fusion and patient graph modelling were combined. Differential privacy introduced a modest reduction compared with the non-private model, but the final performance remained higher than that of all evaluated baselines. Robustness tests further showed that the model can operate when one or more modalities are unavailable. These characteristics make the framework appropriate for heterogeneous hospital environments in which data completeness, disease prevalence and acquisition protocols vary across sites.

The results should nevertheless be interpreted with appropriate caution. They represent a controlled experimental evaluation, and clinical deployment would require external validation on independent hospitals, prospective studies, calibration analysis, subgroup fairness assessment, privacy-budget verification and review by qualified healthcare professionals. Explanations are intended to support clinical judgement rather than serve as autonomous diagnostic evidence.

The proposed FEMT-GAT framework achieved 96.84% accuracy, 96.71% F1-score and 98.42% AUC in the representative evaluation. It outperformed conventional centralized and federated baselines, converged stably, benefited from each major architectural component and maintained useful performance under missing-modality conditions. The global model improved results at every participating hospital, while differential privacy and secure aggregation supported collaborative learning without sharing raw patient data. Multimodal, feature-level and graph-level explanations provided interpretable evidence for the final disease prediction.

5. Conclusion

This paper presented FEMT-GAT, an explainable federated multimodal transformer–graph attention network for privacy-preserving disease prediction. The framework retains sensitive records at their originating hospitals, learns complementary interactions among heterogeneous clinical modalities, and exploits local patient relationships through graph attention. Differential privacy and secure aggregation protect collaborative updates, while multilevel explanation methods provide modality, feature, image, and patient-neighbour evidence. The modular architecture is suitable for a wide range of clinical prediction tasks and offers a practical foundation for transparent and privacy-aware multi-institutional artificial intelligence. Future work can incorporate asynchronous and personalized federation, self-supervised pretraining, domain adaptation across scanners and hospitals, fairness-aware optimization, homomorphic encryption, dynamic temporal patient graphs, uncertainty estimation, and continual learning. Prospective validation with multiple institutions and clinician-centred explanation studies will be necessary to establish safety, usability, generalizability, and regulatory readiness.

References

1. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.
2. H.-Y. Zhou et al., “A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics,” Nature Biomedical Engineering, vol. 7, pp. 743–755, 2023, doi: 10.1038/s41551-023-01045-x.
3. P. Veličković et al., “Graph attention networks,” in Proc. International Conference on Learning Representations (ICLR), 2018.
4. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th International Conference on Artificial Intelligence and Statistics, PMLR, vol. 54, 2017, pp. 1273–1282.
5. M. J. Sheller et al., “Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data,” Scientific Reports, vol. 10, art. no. 12598, 2020, doi: 10.1038/s41598-020-69250-1.
6. I. Dayan et al., “Federated learning for predicting clinical outcomes in patients with COVID-19,” Nature Medicine, vol. 27, pp. 1735–1743, 2021, doi: 10.1038/s41591-021-01506-3.
7. W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, and C. Chen, “A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction,” arXiv:2208.10240, 2022.
8. T. Z. Li et al., “Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine EHRs for pulmonary nodule classification,” arXiv:2304.02836, 2023.
9. N. Rieke et al., “The future of digital health with federated learning,” npj Digital Medicine, vol. 3, art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.
10. M. Adnan et al., “Federated learning and differential privacy for medical image analysis,” Scientific Reports, vol. 12, art. no. 1953, 2022, doi: 10.1038/s41598-022-05539-7.
11. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” Proceedings of Machine Learning and Systems, vol. 2, pp. 429–450, 2020.
12. M. Abadi et al., “Deep learning with differential privacy,” in Proc. ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.
13. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
14. R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE International Conference on Computer Vision, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.