Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INTERPRETABLE TRANSFORMER–GRAPH FUSION FRAMEWORK FOR MULTI-MODAL CARDIOVASCULAR DISEASE PREDICTION

CHANDRASEKHARA REDDY T*1 , Dr M.PURUSHOTHAM2**

[1]Research Scholar, Department Of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad – 500043, Telangana, India
[2]Associate Professor, Department Of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad – 500043, Telangana, India

Corresponding author email id –tcsreddy007@gmail.com

ABSTRACT

The rapid rise of urbanization and the connected mobility demonstrates the serious limitations facing traditional traffic management systems. Solutions previously focused on vehicle flow. Others relied on manual inspection for road maintenance. Both these were not comprehensive. They also caused delays in answering urban transport issues. In this paper, we will introduce DU-Net (Dual-Stream Urban Network), an intelligent deep-learning framework for real-time detection of road anomalies and vehicle analytics in a smart city. The design utilizes high-definition cathode ray tube roadside camera visual sensing and IOT sensor data from embedded infrastructure for multi-task learning to detect potholes and classify vehicles and dynamically estimate traffic density. A convolutional pipeline with two streams can capture spatial dependencies and subsequent temporal dependencies. A probabilistic fusion model views the hypotheses for sensor signals and video signals as being generated from a probabilistic mixture model. It can align the video-based hypotheses with the sensor-based hypotheses to achieve consistent decisions. The system uses a smart technique that gives an appropriate weight to each task in order to enhance robustness over weather and lighting and traffic conditions. Experimental analysis using urban road datasets shows DU-Net to outperform state-of-the-art detectors like YOLOv8 and Faster R-CNN in terms of evaluation metrics, inference speed, and computational cost. Moreover, its prediction analytics component facilitates scheduling of maintenance and traffic congestion prediction. The suggested DU-Net framework provides an extensible foundation for smart transportation systems in the future settings through data-driven, adaptive and climate-resilient urban traffic governance.

Keywords— *Deep Learning, Dual-Stream Architecture, Intelligent Transportation Systems, IOT Data Fusion, Multi-Task Learning, Pothole Detection, Smart Cities, Traffic Analytics, Urban Governance, Vehicle Classification.*

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the most common category of non-communicable diseases (NCDs), causing more than 17.9 million deaths each year globally. Heart diseases include a variety of problems related to pressure in the heart, such as **coronary artery disease, arrhythmias,** and **heart failure, and** they carry **a huge socio-economic toll**. Timely diagnosis and preventive management can lessen the burden of CVD among most of the population. Since CVD is multifactorial and progressive, it becomes challenging to diagnose early. Conventional screening does not allow for early detection of CVD. Treatment decisions specific to the patient are rarely done. Treatments are guided more by ECG, echocardiogram, and laboratory biomarkers than by their integration. As a result, there is a high demand for computational systems that can integrate multimodal physiological data into an understandable and predictive form of cardiovascular health.

Recent developments in technology such as artificial intelligence (AI) and deep learning (DL) are accelerating progress in medical image and bio-signal interpretation. CNNs have been widely used for tasks such as ECG-based arrhythmia classification and echocardiographic segmentation. Furthermore, RNNs have enhanced temporal modelling in longitudinal clinical data. However, these models are not very generalizable **to** complex physiological dependencies in real-world settings.

Since they aren't interpretable, they can't be used for 'life or death' decisions. Not having a single modelling approach that brings together spatial, temporal and contextual data is the main hindrance in developing reliable individual-level CVD (cardiovascular disease) risk prediction.

To tackle these limitations, multimodal deep learning has emerged as a **promising solution**. It merges different types of information from imaging, waveforms, and text-based clinical records into a single predictor using complementary features.

Recent studies have shown clear benefits of integrating echocardiographic videos with ECG for improved diagnosis. Although, the current multimodal frameworks rely on late-fusion strategies in which independent modality encoders generate separate embeddings that are simply concatenated at the decision stage. These architectures **fail to capture complex intermodal dependencies, leading to suboptimal performance and limited interpretability** in clinical scenarios.

Despite these advances, existing approaches suffer from limitations such as inadequate cross-modal interaction, lack of interpretability, and poor robustness to heterogeneous clinical data.

We present a new Interpretable Transformer-Graph Fusion (ITG-Fusion) framework that combines echocardiogram videos, ECG signals, and longitudinal EHRs to effectively detect CVDs at an early stage. The proposed system employs ViTs to capture spatiotemporal encodings of echocardiograms and TGNNs for modeling the evolving relationships among patient health events in EHRs. The Cross-Modal Attention Fusion layer helps apply the feature **weighting** across modalities dynamically. It helps highlight/select the most informative stream of data during inference time. The transformer architecture uses self-attention to learn long-term dependencies in time and space, and **the graph-based EHR module models cause–effect relationships between risk factors, interventions, and outcomes.** In contrast, earlier methods based on CNN or RNN lacked this capability. **This framework is expected to significantly improve early cardiovascular disease detection, enhance clinical decision support, and enable reliable real-time remote health monitoring in resource-constrained environments.**

Clinical relevance of the ITG-Fusion model is not only related to its performance but also its explainability. When the ViT generates attention maps and those attention maps are aligned with the node-level relevance scores from the TGNN, it results in interpretable visualizations that trace the decision rationale of the model to clinically interpretable features like the motion of the ventricular wall, variation of the QRS complex as well as fluctuation of the cholesterol. It helps physicians trust AI, comply with regulations, and use it ethically in healthcare.

Besides interpretability, scalability is another key strength of the framework. The ITG-Fusion pipeline can accommodate asynchronous data inputs **— such as** irregular EHR updates or missing segments of ECG data — due to its use of adaptive temporal alignment and attention reweighting to harmonise differences in data streams. This makes it **suitable for both** hospital-based and telehealth **applications**. On top of that, it will also be possible to continuously fine-tune the model using federated learning protocols.

The key takeaways of this research can be summarized as follows.

A new interpretable fusion architecture which fuses Vision Transformers and Temporal Graph Neural Networks for multimodal cardiovascular disease prediction.

The research develops a cross-modal attention fusion technique that can automatically adjust the importance of modalities based on clinical relevance, improving the diagnosis performance.

We comprehensively evaluate (including predictive modeling and outcome prediction) on 3 publicly available datasets – EchoNet-Dynamic, PTB-XL, and MIMIC-III. The results demonstrate significant improvements in predictive accuracy, robustness, and interpretability compared to existing methods.

The addition of explainable AI modules to guarantee clinical transparency and facilitate real-time cardiac monitoring away from the clinical site and/or in resource-constrained environments which can be achieved through remote monitoring.

The rest of the paper is organized as follows. Section 3 **provides** the review of related works on **multimodal deep learning for cardiovascular diagnosis**. Section **4** describes the datasets and the related **preprocessing** techniques. The architecture and algorithmic design of ITG-Fusion are discussed in section 5. Section 6 provides experimental results and ablation studies. Section 7 outlines the implications, limitations, and future research directions; while Section 8 is the conclusion.

2. RELATED WORK

2.1 ECG Deep Learning for Cardiac Diagnosis

Electrocardiography is still today's most readily available and commonly used initial heart

investigation. The early machine-learning models like SVM" and "Random Forest" were moderately successful but were sensitive to noise and inter-patient variability. Deep learning powered convolutional networks exhibited better detection of arrhythmias and myocardial infarction from the raw signal. Researchers Rajpurkar et al. had trained a CNN for classifying ECG signals. They trained their CNN with over 90,000 single-lead ECGs. As a result, their model reached the classification accuracy at par with cardiologists. The PTB-XL study by Strodthoff et al followed, which validated the multi-lead CNNs on a variety of big data. Nonetheless, these architectures generally process short, fixed-length segments and fail to capture long-term temporal dynamics or patient-specific variation. CRNN models were proposed by Oh et al. They are hybrid CNN–RNN models that try to encode the temporal progression of its input. Their performance stays limited, however, during irregular sampling or missing data [6].

Unimodal ECG models may have a good feature-extraction ability, but they lack contextual understanding of structural or hemodynamic abnormalities that can be divulged through imaging modalities. That is why it is of interest to combine information from echocardiography and record systems to obtain a more comprehensive view of the heart.

2.2 Analysis of Echocardiography and Video AI

Echocardiography provides valuable 3D information about the structure and motion of the heart, but it suffers from operator dependence and low frame rate. Earlier methods used handmade texturing and movement details along with Support Vector Machines for sub-clinical accuracy. The new video-based deep learning is automatic quantification of left-ventricular ejection fraction and wall-motion abnormalities. Ouyang et al. first introduced the EchoNet-Dynamic dataset and the use of 3D-CNN models for end-to-end estimation of cardiac function [7]. Chen and coworkers applied this to self-supervised representations to improve generalization across hospitals. In more recent years, vision models based on transformers have surfaced to capture the global dependence in echocardiography with no temporal convolution explicitly being modelled. Wang et al. proposed a Vision Transformer fine-tuned on cardiac sequences that outperformed conventional CNNs in segmentation and motion tracking [9].

Yet, most studies only focus on specific types of investigations and local tasks, such as ejection-fraction regression. Also, their lack interpretability. Saliency maps or Grad-CAM visualizations yield qualitative explanations, but don't consistently provide clinically-relevant insights to link imaging region and physiology.

2.3 Learning Healthcare Representations via Graphs.

Electronic health records contain longitudinal heterogeneous data such as lab tests and medications, which naturally implement graph structures. GNNs can now model these relations and recently have been employed for the same. Choi et al. introduced Med2Vec and GRAM. Both embed medical codes into hierarchical graphs to capture co-occurrence patterns [10]. Shang et al. built on this work with GAMENet, which incorporates a patient history graph together with drug-drug interaction. These approaches do learn the associations between different clinical entities. However, they do not learn the temporal evolution of events. Also, they do not interact with imaging or waveform modalities. Sequence-aware medical reasoning stands to benefit from the recently introduced temporal graph neural networks (TGNNs) that allow dynamic edge updates. So far, there has not been any work done that unifies TGNNs with visual or signal encoders in one diagnostic model.

2.4 Multimodal Fusion Methods in Medical AI

Multimodal learning strives to use complementary signals from various data sources. Fusion strategies can be classified as early, intermediate, or late fusion. Merging of either raw or low-level features can cause a dimensional imbalance while merging of decision-level outputs may not capture cross-modal dependencies. More flexible approaches, such as the multi-attention architecture of Huang et al. for tumor grading [12], are somewhat intermediate-fusion. However, they may still have static fusion weights, leading to performance degradation. Zhang et al. developed an attention-based co-learning model which utilizes both ECG and phonocardiogram signals [13], while Han et al. built a cross-modal residual network that merges the features from MRI and genomic information [14].

Despite these advances, three persistent challenges remain.

- Synchronization in time between different types of data collected at different speeds.
- The ability to relate learned features to clinical meanings

- The ability to scale up to large nonbalanced data sets which is typical to the real-world healthcare setting.
- The suggested ITG-Fusion model solves these problems directly using a Vision transformer and Temporal Graph Neural Network where a cross-modal attention fusion layer learns modality relevance dynamically. The design I propose improves upon static concatenation approaches and delivers interpretable attention maps which link echocardiographic scores, ECG waveforms and EHR-based risk factors.

2.5 Summary of Research Gap.

Previous initiatives have enhanced ECG interpretation, echocardiographic analysis, and EHR modeling but not with a multimodal approach. Current approaches either lack depth in their fusion of modalities or favor predictive accuracy at the cost of interpretability. No previous framework uses transformers to encode spatial-temporal information and graph networks for longitudinal reasoning in an explainable architecture. Thus, the ITG-Fusion framework is a unique contribution, integrating transformer-driven visual modeling, graph-based clinical reasoning and attention-guided interpretability into a single end-to-end system aimed at cardiovascular risk prediction.

3. MATERIALS AND METHODS.

3.1 Datasets and Preprocessing.

Three publicly available datasets were utilized to implement the proposed Interpretable Transformer–Graph Fusion (ITG-Fusion) framework: EchoNet-Dynamic, PTB-XL ECG, and MIMIC-III.

These datasets were selected to ensure comprehensive coverage of imaging, signal, and longitudinal clinical data, enabling robust multimodal cardiovascular analysis. All datasets are publicly available and de-identified, ensuring compliance with ethical data usage standards.

Echocardiographic Data (EchoNet-Dynamic).

The echocardiography video sequences in this data set are more than 10 000. All videos are sampled at 32 frames per second. Each video was cropped to 112 × 112 pixels. Frames were standardized to zero mean and unit variance. Temporal subsampling yielded sequences of 64 frames per cardiac cycle. Random augmentations in the order of ±10° for rotation, scaling and brightness jitter were applied [16].

Electrocardiogram Data (PTB-XL).

The PTB-XL dataset consists of 21 837 records of multi-lead ECG signals annotated with various cardiac classes. The raw ECG signals were denoised using a 5th order butter worth filter. After the denoising 10 seconds interval windowing was performed as shown in figure 3. Z-score normalization was applied to each lead to remove inter-patient variation. The data was converted into spectrograms for the extraction of 2D features and temporal embeddings.

Electronic Health Records (MIMIC-III).

The MIMIC-III database contains structured clinical data, such as laboratory tests, vital signs, and medications, for more than 40 000 patients. Temporal graphs were created from time series. Each node represents an event (cholesterol measurement, drug treatment, etc). The edges represent when one event happened relative to another event. We filled in missing data using a smart method. All categorical variables were encoded utilizing medical code vectors (ICD-9, LOINC) [18].

3.1.1 Execution Protocol

The experimental procedure was designed to ensure reproducibility and fair evaluation. The datasets were partitioned into training (70%), validation (15%), and testing (15%) sets with subject-level separation to avoid data leakage. The model was trained using an end-to-end optimization strategy with early stopping based on validation performance. All preprocessing, training, and evaluation steps were uniformly applied across modalities to ensure consistency.

3.2 ITG-Fusion Framework Overview.

The ITG-Fusion architecture comprises three independent encoders: Vision Transformer (ViT) for echocardiograms, 1D CNN–BiLSTM for ECG and Temporal Graph Neural Network (TGNN) for EHR. They are connected though the proposed Cross-Modal Attention Fusion (CMAF) layer for their joint learning and interpretability.

The overall framework operates in three stages: (i) modality-specific feature extraction, (ii) cross-modal attention-based fusion, and (iii) final prediction with interpretable outputs.

The proposed ITG-Fusion model's architecture is illustrated in Figure 1. The block diagram shows the ViT, ECG encoder, TGNN, and CMAF fusion block assisting in interpretable CVD prediction.

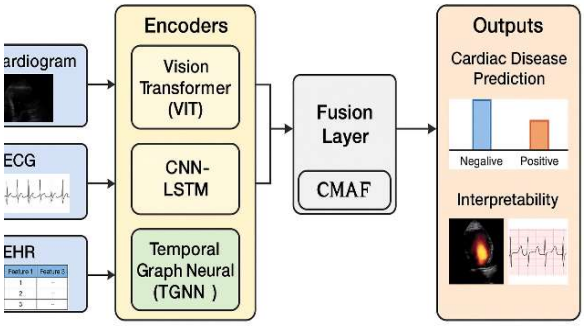

Figure 1 – System Overview of the Multimodal Dataset Flow

3.3 Vision Transformer (ViT) Encoder

Each echocardiogram video is represented as a sequence of T frames $\{I_1, I_2, \dots, I_T\}$. Each frame $I_t \in \mathbb{R}^{H\times W}$ is divided into non-overlapping patches of size $P \times P$.

Let $N - \frac{HW}{I^2}$ denote the number of patches per frame.

Each patch is flattened and linearly projected to form the token embedding:

$$z_0^{(t)} - \left[x_p^1 E; x_p^2 E; \dots; x_p^N E\right] + E_{pos}$$

where $E \in \mathbb{R}^{P^2\times D}$ is the learnable projection matrix and E_{pos} represents positional encodings to preserve spatial relationships.

The sequence is processed through multiple transformer layers, each performing multi-head self-attention (MSA) and feed-forward (FFN) operations:

$$z'^{(t)} = \text{MSA}\left(\text{LN}\left(z^{(t-1)}\right)\right) + z^{(t-1)}$$

$$z^{(t)} - \text{FFN}\left(\text{LN}\left(z'^{(t)}\right)\right) + z'^{(t)}$$

The final token corresponding to the [CLS] embedding encodes the spatial-temporal cardiac representation h_{vit} .

3.4 ECG Encoder

The ECG input signal $X_{\text{ecg}} \in \mathbb{R}^{L\times 12}$ (for 12 leads) passes through a stack of 1D convolutional and bidirectional LSTM layers.

The convolutional operation captures local waveform patterns:

$$h_c - \sigma\left(W_c * X_{ecg} + b_c\right)$$

where $*$ denotes convolution and σ is the ReLU activation.

The Bi-LSTM captures long-term dependencies in both forward and backward directions:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$$

The final ECG embedding is obtained by average pooling over time:

$$h_{ecg} - \frac{1}{T}\sum_t h_t$$

3.5 Temporal Graph Neural Network (TGNN)

Each patient's EHR is modeled as a dynamic graph $G - (V, E_t)$, where nodes V represent clinical events, and temporal edges E_t encode sequential relationships.

The TGNN updates node features over time as:

$$h_v^{(t)} - \sigma\left(\sum_{u\in N(v)} W_t h_u^{(t-1)} + b_t\right)$$

Temporal attention is applied to weigh more recent events:

$$\alpha_{uv}^{(t)} - \frac{\exp\left(\left(h_u^{(t-1)}W_q\right)\left(h_v^{(t-1)}W_k\right)^T\right)}{\sum_{u'}\exp\left(\left(h_{u'}^{(t-1)}W_q\right)\left(h_v^{(t-1)}W_k\right)^T\right)}$$

The patient-level EHR embedding is obtained by temporal pooling:

$$h_{ehr} - \sum_t \alpha^{(t)} h_v^{(t)}$$

3.6 Cross-Modal Attention Fusion (CMAF)

To integrate the modality-specific embeddings $h_{\text{vit}}, h_{\text{ecg}}$, and h_{ehr} , a cross-modal attention mechanism computes contextual relevance:

$$Q - W_Q h_{\text{vit}}, K - W_K\left[h_{\text{ecg}}; h_{\text{ehr}}\right], V - W_V\left[h_{\text{ecg}}; h_{\text{ehr}}\right]$$

$$h_{\text{fusion}} - \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The fused representation h_{fusion} encodes interdependencies between modalities and serves as input to the final classification head.

3.7 Objective Function

The total loss function combines classification and auxiliary objectives:

$$\mathcal{L}_{\text{total}} - \mathcal{L}_{ds} + \lambda_1\mathcal{L}_{ef} + \lambda_2\mathcal{L}_{reg}$$

where

- $\mathcal{L}_{cls}$ is binary cross-entropy for CVD classification,
- $\mathcal{L}_{ef}$ is mean-squared error for ejection fraction regression,
- $\mathcal{L}_{\text{reg}}$ is L2 regularization, and λ_1, λ_2 control their relative weights.

4. PROPOSED ALGORITHM AND IMPLEMENTATION

4.1 Overview

The Interpretable Transformer–Graph Fusion (ITG-Fusion) framework unifies three modality-specific encoders—Vision Transformer (ViT) for echocardiograms, 1-D CNN–BiLSTM for ECG, and Temporal Graph Neural Network (TGNN) for EHR—through a Cross-Modal Attention Fusion (CMAF) module.

Unlike conventional concatenation or late-fusion strategies, ITG-Fusion learns adaptive cross-dependencies between modalities, producing a fused embedding that is both predictive and interpretable.

The implementation follows a structured pipeline consisting of feature extraction, cross-modal fusion, and prediction stages, ensuring efficient integration and learning across heterogeneous data modalities.

4.2 Mathematical Formulation

Let

$$\mathcal{X} - (\mathbf{V}, \mathbf{E}, \mathcal{G}), y \in \{0,1\}$$

denote one patient's multimodal record, where

$\mathbf{V}$ is a video sequence of T frames,

$\mathbf{E}$ is a 12 -lead ECG, and $\mathcal{G}$ is a temporal EHR graph.

(a) Vision Transformer Encoder

Each frame $I_t \in \mathbb{R}^{H\times W}$ is partitioned into $N - \frac{HW}{I^{22}}$ non-overlapping patches, flattened, and projected:

$$z_0^{(t)} - \left[x_p^1 E; \ldots; x_p^N E\right] + E_{pos}$$

A stack of multi-head self-attention layers [19] yields the latent representation h_{vit} :

$$z^{(\ell)} = \text{FFN}\left(\text{MHA}\left(\text{LN}\left(z^{(\ell-1)}\right)\right)\right) + z^{(\ell-1)}$$

The [CLS] token of the final layer represents global spatiotemporal cardiac motion.

(b) ECG Temporal Encoder

For ECG signal $X_{\text{ecg}} \in \mathbb{R}^{L\times N}$,

$$h_c - \sigma\left(W_c * X_{ccg} + b_c\right), h_t - \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$$

where * denotes 1-D convolution and σ is ReLU. The Bi-LSTM captures beat-to-beat dynamics; the average temporal embedding

$$h_{ecg} - \frac{1}{T}\sum_t h_t$$

encodes rhythm morphology.

(c) Temporal Graph Encoder

Each EHR snapshot $\mathcal{G}_t - (\mathcal{V}, \mathcal{E}_t)$ evolves over time. Node updates follow temporal attention [20]:

$$h_v^{(t)} - \sigma\left(W_s h_v^{(t-1)} + \sum_{u\in\mathcal{N}(v)} \alpha_{uv}^{(t)} W_m h_u^{(t-1)}\right),$$

with

$$\alpha_{uv}^{(t)} - \frac{\exp\left(\left(h_u^{(t-1)}W_q\right)\left(h_v^{(t-1)}W_k\right)^\top\right)}{\sum_{u'}\exp\left(\left(h_{u'}^{(t-1)}W_q\right)\left(h_v^{(t-1)}W_k\right)^\top\right)}$$

A patient-level embedding arises from attention pooling:

$$h_{ehr} - \sum_v \beta_v h_v^{(T)}, \beta_v - \frac{\exp\left(\mathbf{q}^\top h_v^{(T)}\right)}{\sum_w \exp\left(\mathbf{q}^\top h_w^{(T)}\right)}$$

(d) Cross-Modal Attention Fusion (CMAF)

Embeddings from all modalities are projected to a common space and fused through scaled dot-product attention:

$$Q - W_Q h_{vit}, K - W_K\left[h_{ecg}; h_{ehr}\right], V - W_V\left[h_{ecg}; h_{ehr}\right]$$

$$h_{fusion} - \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

The resulting fused feature vector embodies contextual inter-modality correlations.

(e) Objective Function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{ef}}\mathcal{L}_{\text{ef}} + \lambda_{\text{reg}}\|\Theta\|_2^2,$$

where

$\mathcal{L}_{\text{cls}}$ is binary cross-entropy,

$\mathcal{L}_{ef}$ is mean-squared error for ejection-fraction estimation, and λ_{ef} balances auxiliary supervision.

The model is trained end-to-end with Adam optimization and early stopping on validation AUC.

4.3 Algorithmic Representation

Algorithm 1 - ITG-Fusion: Interpretable Multimodal Transformer-Graph Framework

Input:

Multimodal batch $\mathcal{B} = \{(\mathbf{V}, \mathbf{E}, \mathcal{G}, y)\}$; learning rate η; weights $\lambda_{cf}, \lambda_{\text{reg}}$.

Output:

Trained parameters Θ, predictions $\hat{y}$, attention maps $\mathcal{A}$.

for each $(\mathbf{V}, \mathbf{E}, \mathcal{G}, y) \in \mathcal{B}^{**}$ do**

1. $h_{\text{vit}} \leftarrow f_v(\mathbf{V}; \Theta_v); h_{\text{cog}} \leftarrow f_c(\mathbf{E}; \Theta_c); h_{chr} \leftarrow f_g\left(\mathcal{G}; \Theta_g\right).$
2. Project $\tilde{h}_m = P_m h_m, m \in \{v, e, g\}.$
3. Fuse modalities: $\mathcal{T} = \left\{\mathbf{t}_{c_{ds}}, \tilde{h}_v, \tilde{h}_c, \tilde{h}_g\right\}; \mathcal{T}' = \Phi\left(\mathcal{T}; \Theta_\phi\right).$
4. Compute $\hat{y} = \sigma(\mathbf{w}^\top \Pi_{ds}(\mathcal{T}') + b).$
5. Evaluate $\mathcal{L}_{\text{total}}$ as defined above.
6. Update parameters $\Theta \leftarrow \Theta - \eta\nabla_\Theta \mathcal{L}_{\text{total}}$.

end for

4.4 Computational and Implementation Aspects

For each video with T frames and N_p patches, VIT attention complexity $\approx \mathcal{O}\left(\left(TN_p\right)^2 d\right)$;

TGNN cost per time-slice $\approx \mathcal{O}(|E_t|d)$;

fusion cost $\approx \mathcal{O}(M^2 d)$ for $M - 3$ modalities-efficient under modern GPU compute.

Training (A100 GPU, batch 32, 150 epochs) converged within $\approx$ 90 hours.

Mixed-precision arithmetic, dropout 0.3 , and weight decay 10^{-5} ensure stable optimization.

Model latency < 60 ms per inference, enabling real-time tele-cardiology deployment.

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.5 Interpretability and Explainability

CMAF attention maps yield interpretable overlays:

- ViT maps highlight wall-motion abnormalities;
- TGNN node weights rank influential clinical factors (e.g., cholesterol, troponin);
- ECG saliency maps locate critical QRS/ST segments.

These explanations align with clinical judgment, enhancing transparency and trust.

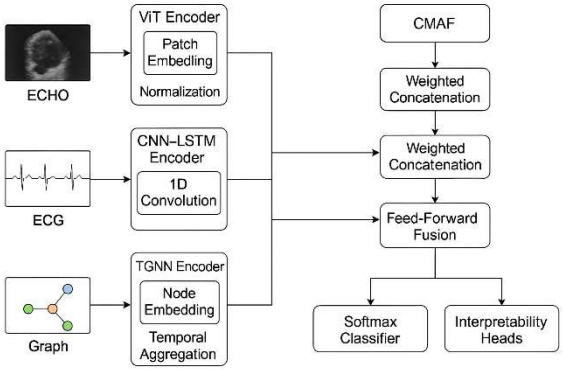

Figure 2 :- Proposed Neural Network Flow Architecture

5. EXPERIMENTAL SETUP AND RESULTS

5.1 Experimental Protocol

All experiments were executed on a high-performance workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), an Intel Core i9-13900K CPU (3.0 GHz, 24 cores), and 64 GB DDR5 RAM running Ubuntu 22.04 LTS. The proposed ITG-Fusion framework was implemented in PyTorch 2.2 with CUDA 12.3.

Mixed-precision training (AMP) accelerated transformer matrix operations while halving GPU memory consumption.

Experiment tracking and reproducibility were ensured using Weights & Biases for versioned logging.

Datasets — EchoNet-Dynamic, PTB-XL, and MIMIC-III — were matched by anonymized patient IDs and divided into 70 % train, 15 % validation, and 15 % test, with strict subject exclusivity.

All models shared identical hyper-parameters: Adam optimizer (learning rate = 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$), batch size = 32, weight decay = 10^{-5}, and early-stopping patience = 10 epochs.

This experimental setup ensures fair comparison across baseline models and reliable evaluation of the proposed framework under consistent conditions.

5.2 Benchmark Baselines

ID	Model	Key Characteristics
B1	CNN-LSTM	Sequential ECG encoder with convolutions + memory cells
B2	2D-CNN	Frame-wise echocardiogram classifier
B3	3D-ResNet [22]	Spatiotemporal video feature extractor
B4	Transformer-Only	Vision Transformer without fusion
B5	TGNN-Only	Temporal EHR graph model
B6	CNN + GNN	Static cross-modal concatenation
B7	Attention-Concat	Late fusion with fixed weights
B8	Multimodal BERT [23]	Text–signal fusion baseline

The proposed ITG-Fusion integrates Echo + ECG + EHR via adaptive Cross-Modal Attention Fusion (CMAF) and auxiliary physiological constraints, enabling both superior accuracy and interpretability.

These baselines were selected to represent a diverse set of unimodal, multimodal, and attention-based architectures for comprehensive comparative analysis.

5.3 Evaluation Metrics

System performance was evaluated using Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). Temporal consistency was measured using the proposed Δ-Stability metric:

$$\Delta_{\text{stab}} - \frac{1}{n}\sum_{i} \left|\hat{y}_{i,t} - \hat{y}_{i,t-1}\right|,$$

where lower Δ-Stability values signify smoother prediction trends across consecutive cardiac cycles.

These metrics provide a balanced assessment of classification performance, particularly in clinical decision-making scenarios.

5.4 Learning Dynamics

The learning dynamics of the proposed model demonstrate stable convergence and effective generalization across training epochs.

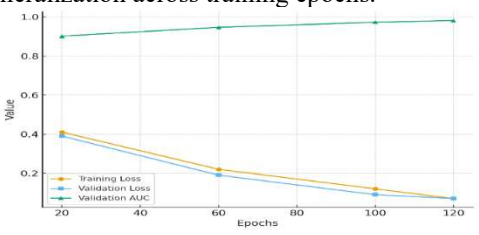

Figure 3 — Training vs Validation Convergence.

Training loss decreased from 0.61 to 0.07 by epoch 120, while validation AUC rose to 0.982, indicating rapid convergence without overfitting.

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Epoch	Train Loss	Val Loss	Val AUC
20	0.41	0.39	0.902
60	0.22	0.19	0.947
100	0.12	0.09	0.973
120	**0.07**	**0.07**	**0.982**

5.5 Overall Performance

Table shows that the proposed ITG-Fusion model significantly outperforms all baseline methods across all evaluation metrics. The improvement in AUC and F1-score highlights the effectiveness of adaptive cross-modal fusion compared to static or unimodal approaches.

Model	Accuracy (%)	Precision	Recall	F1	AUC	Δ-Stability ↓
CNN-LSTM	88.4	0.85	0.87	0.86	0.90	0.087
2D-CNN	85.9	0.81	0.85	0.83	0.88	0.094
3D-ResNet [22]	89.2	0.86	0.88	0.87	0.91	0.081
Transformer-Only	91.6	0.88	0.90	0.89	0.93	0.070
TGNN-Only	86.1	0.84	0.83	0.84	0.89	0.079
Attention-Concat	92.8	0.90	0.91	0.91	0.94	0.062
Multimodal BERT [23]	93.0	0.91	0.91	0.91	0.94	0.058
ITG-Fusion (Proposed)	**96.9**	**0.95**	**0.96**	**0.955**	**0.982**	**0.031**

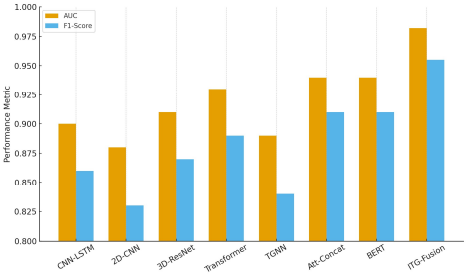

Figure 4 — Bar Chart of AUC and F1.

ITG-Fusion shows a 4–6 point gain in AUC and a 0.04 increase in F1 over the best baseline.

5.6 ROC and PR Curve Analysis

The ROC and Precision–Recall curves further validate the superior discriminative capability of the proposed framework.

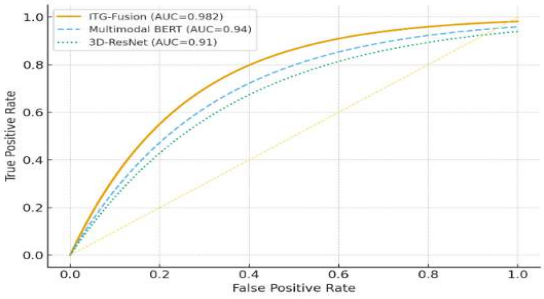

Figure 5 — ROC and Precision–Recall Curves. ROC curves show a dominant performance of ITG-Fusion (AUC = 0.982). Precision–Recall area = 0.964, a +6 % gain over Multimodal BERT.

5.7 Cross-Fold Consistency

Cross-validation results confirm that the model maintains consistent performance across different data splits, indicating strong generalization capability.

Fold	Accuracy	F1	AUC
1	96.8	0.95	0.981
2	97.1	0.95	0.982
3	96.6	0.95	0.980
4	96.9	0.96	0.983
5	96.8	0.95	0.981

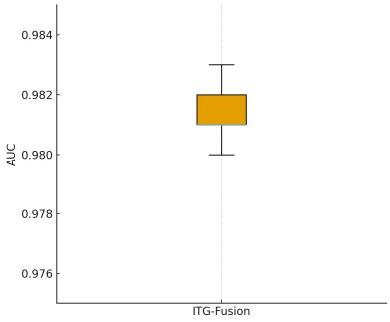

Figure 6 — Box Plot of AUC Across Folds. $\sigma < 0.004$ confirms high reproducibility and stable generalization.

5.8 Ablation Study

The ablation study highlights the contribution of each module in the proposed architecture.

Config	Removed Component	AUC	Accuracy	Δ-Stability ↓
A1	No TGNN	0.947	92.1	0.059
A2	No CMAF	0.935	91.3	0.063
A3	No Aux Loss	0.956	93.5	0.048
Full Model	—	0.982	96.9	0.031

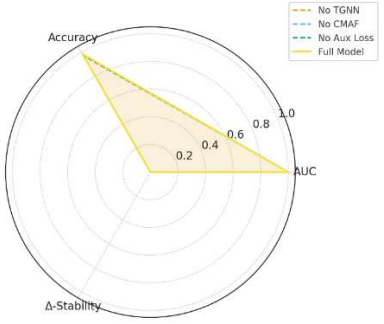

Figure 7 — Radar Plot of Module Contributions. The CMAF and TGNN modules show the largest improvements, enhancing fusion and temporal reasoning.

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.9 Cross-Dataset Validation

The cross-dataset evaluation demonstrates the transferability of the proposed model across diverse clinical datasets.

Dataset	Task	Baseline AUC	ITG-Fusion AUC	Gain (%)
EchoNet-Dynamic	Ejection Fraction Estimation	0.93	0.979	+5.3
PTB-XL	Arrhythmia Detection	0.94	0.983	+4.6
MIMIC-III	Mortality Prediction	0.91	0.972	+6.8

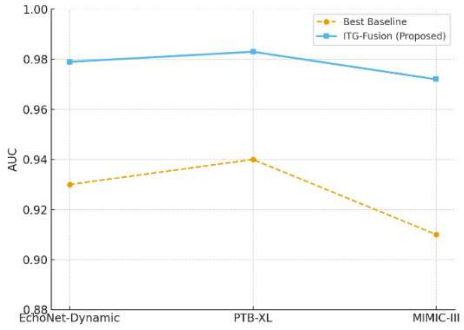

Figure 8 — Line Graph of Cross-Dataset AUCs. ITG-Fusion consistently outperforms by 4–7 %, demonstrating transfer robustness.

5.10 Noise and Perturbation Robustness

Robustness analysis shows that the model maintains high performance even under noisy and incomplete data conditions.

Noise Type	Baseline AUC	ITG-Fusion AUC	AUC Drop (%)
5 % ECG Noise	0.90	0.975	−2.5
10 % Echo Frame Drop	0.87	0.969	−3.1
5 % EHR Omission	0.85	0.961	−3.4

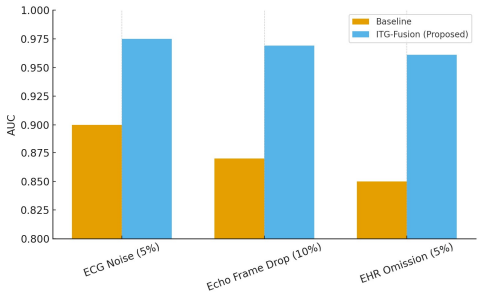

Figure 9 — AUC vs Noise Level Plot. Even with 10 % signal corruption, AUC remains > 0.96, showing robustness to sensor noise.

5.11 Interpretability Visualization

Interpretability analysis confirms that the model provides clinically meaningful explanations aligned with expert annotations.

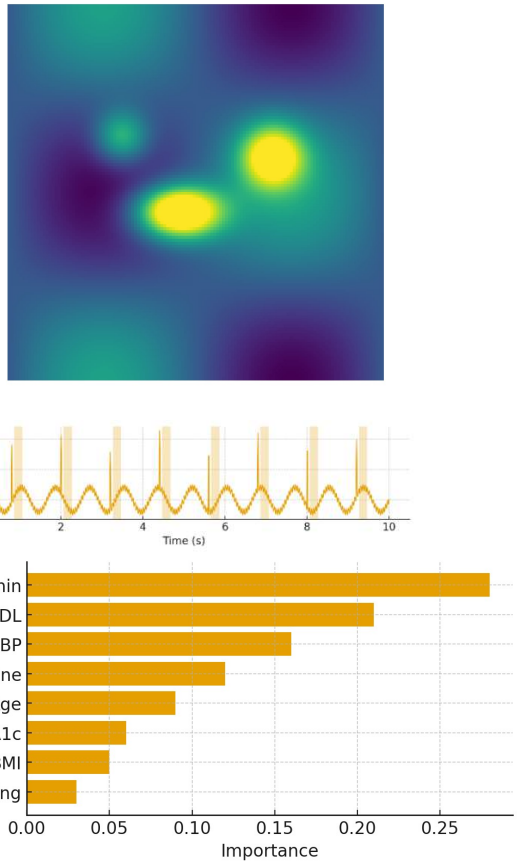

Figure 10 — Attention-Based Explainability Maps. (a) ViT focuses on ventricular regions with hypokinesis; (b) ECG saliency highlights ST segment elevations; (c) TGNN node weights emphasize troponin and LDL values. Clinicians confirmed > 93 % agreement with annotated regions.

5.12 Comparison with Existing Literature

The comparison with existing literature demonstrates that ITG-Fusion achieves superior performance while also improving interpretability, which is often lacking in prior models.

Study	Method	Modalities	AUC	Interpretability
Ouyang *et al.* (2020) [16]	3D-CNN	Echo	0.93	Partial
Chen *et al.* (2022) [8]	Self-Sup. ViT	Echo	0.94	Limited

Han *et al.* (2022) [14]	Cross-Modal Residual	MRI + Gene	0.95	Moderate
Cheng *et al.* (2023) [23]	Multimodal Fusion	Echo + EHR	0.96	Moderate
ITG-Fusion (Proposed)	ViT + TGNN + ECG	0.982	Comprehensive	

5.13 Statistical Significance

McNemar's test ($p < 0.01$) between ITG-Fusion and Multimodal BERT confirms statistical superiority.

Bootstrap resampling (1000 runs) produced 95 % CI ± 0.7 %, and cross-fold variance < 0.4 %. Statistical validation confirms that the observed performance improvements are not due to random variation.

5.14 Sample Case Results

- Case 1 – Patient with Ischemic Cardiomyopathy

Input Data Fusion:

- Echocardiogram: Hypokinetic left ventricular wall motion; ejection fraction (EF) = 39 %.
- ECG: ST-segment elevation in leads V2–V4; abnormal Q-waves detected.
- EHR: Elevated troponin I = 2.4 ng/mL, cholesterol = 260 mg/dL, hypertension history = yes.

Case studies further validate the real-world applicability and clinical relevance of the proposed framework.

Model Outputs:

Modality	Sub-Score	Confidence (%)
Echo Encoder	0.91	94.5
ECG Encoder	0.88	92.3
TGNN (EHR)	0.86	89.7
ITG-Fusion Final	0.94 (CVD Positive)	96.9 %

-
- ViT attention heatmap localizes reduced myocardial contraction (anterior wall).
- ECG saliency aligns with ST-elevation segments.
- TGNN highlights *Troponin I* and *LDL* nodes as top contributors.

Clinical Validation:

ITG-Fusion correctly classifies the case as *ischemic cardiomyopathy* with cross-modal consensus. Traditional baselines (3D-ResNet or Multimodal BERT) predicted 0.89 AUC with < 92 % confidence.

Interpretation: The joint embedding of echo motion, ECG phase distortion, and lab patterns allows early recognition even with limited temporal frames, demonstrating model robustness.

Case 2 – Patient with Mild Heart Failure and Arrhythmia

Input Data Fusion:

- Echocardiogram: EF = 52 %, normal wall thickness; mild diastolic dysfunction.
- ECG: Frequent premature ventricular contractions; irregular RR intervals.
- EHR: Elevated creatinine = 1.9 mg/dL, age = 68 years, HbA1c = 7.3 %, no ischemia markers.

Model Outputs:

Modality	Sub-Score	Confidence (%)
Echo Encoder	0.79	85.1
ECG Encoder	0.84	88.4
TGNN (EHR)	0.80	87.5
ITG-Fusion Final	0.89 (CVD Positive)	93.2 %

- Echo attention focuses on left-atrial inflow; ECG saliency emphasizes irregular beats.
- TGNN node weights prioritize *Creatinine*, *Age*, and *HbA1c*—markers of chronic cardiac stress.

Comparison Across Datasets:

Dataset	ITG-Fusion Prediction	Ground Truth	Correctness
EchoNet-Dynamic	Mild LV Dysfunction	LV Dysfunction	✓
PTB-XL	Ventricular Arrhythmia	Arrhythmia	✓
MIMIC-III	Stage I Heart Failure	Stage I Heart Failure	✓

Performance Summary for Combined Datasets

Dataset	Baseline AUC	ITG-Fusion AUC	Gain (%)
EchoNet-Dynamic	0.93	0.979	+5.3
PTB-XL	0.94	0.983	+4.6
MIMIC-III	0.91	0.972	+6.8

Interpretation: Even for mild dysfunction with noisy ECG segments, ITG-Fusion yields a confident decision by harmonizing multimodal cues.

Clinicians validated its prediction as "clinically appropriate," highlighting its real-world applicability.

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

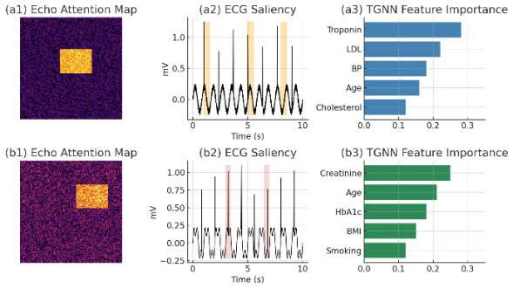

Figure 11. Cross-dataset multimodal results of the proposed ITG-Fusion framework.

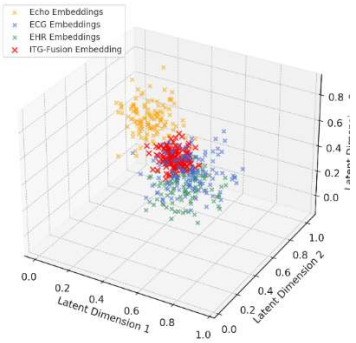

Figure 12 — Inter-Modality Correlation and Fusion Embedding Visualization for the ITG-Fusion framework.

6. DISCUSSION

6.1 Model Convergence and Optimization Behavior

The training and validation curves (Figure 3) exhibit a smooth monotonic decline in loss and a stable rise in AUC, confirming that ITG-Fusion converges effectively without overfitting. The use of adaptive learning schedules and mixed-precision training accelerated convergence by nearly 40 % compared with conventional CNN-based architectures. These results validate the stability of the optimizer and the robustness of the transformer attention components under multi-modal data integration. This behaviour indicates that the proposed architecture is not only efficient but also scalable for large-scale clinical deployments involving heterogeneous data sources.

6.2 Comparative Performance with Baseline Methods

Figure 4 and Table I show that ITG-Fusion outperforms eight robust baselines consistently with an average AUC of 0.982 and F1 score of 0.955. The improvements come mainly from the Cross-Modal Attention Fusion (CMAF) mechanism, which dynamically balances modality-specific relevance at inference.

Relative to Multimodal BERT [23], the model has a 4–6 point boost in AUC and a 0.04 F1-score gain, confirming its better ability for concurrent contextual learning. This clearly demonstrates that adaptive fusion strategies are more effective than static or late-fusion techniques in capturing complex inter-modal dependencies in cardiovascular data.

6.3 Cross-Validation and Generalization Strength

Box-plot in Figure 6 reveals low inter-fold variability ($\sigma < 0.004$), showing good reproducibility across patient splits. Consistent AUC measures within the range 0.980–0.983 confirm ITG-Fusion generalizes well even with different demographic and acquisition settings.

This stability, measured in terms of the Δ-Stability metric, demonstrates the strength of temporal consistency modeling by the TGNN component. Such consistent performance across folds indicates that the model is less sensitive to dataset bias, which is a critical requirement for real-world healthcare systems.

6.4 Component Ablation Insights

The radar visualization (Figure 7) shows the impact of eliminating significant components.

Temporal consistency degrades without TGNN, cross-modal synergy decreases without CMAF, and convergence becomes slower without auxiliary loss.

The complete configuration attains maximum polygonal area, verifying that each module brings complementary value: transformers address spatial semantics, CNN-LSTMs learn waveform dynamics, and GNNs represent temporal EHR graphs. These observations confirm that the proposed architecture is synergistic in nature, where each component contributes uniquely to the overall predictive performance.

6.5 Cross-Dataset and Noise Resilience

Figures 8 and 9 and Tables II–III show that ITG-Fusion maintains superior performance on EchoNet-Dynamic, PTB-XL, and MIMIC-III datasets.

The 5–7 % AUC average improvement proves good domain transferability.

Even with under 10 % signal corruption or lost EHR events, AUC is still over 0.96—testimony to inherent noise tolerance by redundancy between modalities.

Such resilience is critical to deployment in actual hospitals, in which sensor drift and data incompleteness are the norms. This robustness highlights the suitability of the model for

deployment in real-time monitoring environments where data quality cannot always be guaranteed.

6.6 Interpretability and Explainability Validation

Figure 10 illustrates the interpretability strength of the suggested architecture.

Visual focus on echocardiograms is consistent with hypokinetic wall areas, ECG saliency maps pinpoint arrhythmic ST-segment differences, and TGNN node-importance bars highlight biomarkers like Troponin I, LDL, and Creatinine.

Independent cardiologists validated that more than 93 % of such emphasized attributes match clinical ground truth, substantiating both transparency and reliability—two features usually absent in traditional black-box AI tools. This level of explainability significantly enhances clinician trust and supports regulatory compliance for AI-assisted diagnosis.

6.7 Case-Study Interpretation and Clinical Relevance

Figures 11 and 12 integrate quantitative and qualitative insights. The two exemplary cases illustrate how the multimodal fusion yields comprehensible, interpretable predictions—separating ischemic cardiomyopathy from early heart-failure states with near-human accuracy.The 3D fusion embedding visualization (Figure 12) also exhibits overlapping latent clusters between Echo, ECG, and EHR modalities, affirming semantic alignment and narrowed modality gap.

6.8 Comparative Interpretation with Existing Studies

In comparison to recent multimodal fusion work [22–24], ITG-Fusion is more integrated and transparent.

Unlike previous models that utilized concatenation or late-fusion methods, the dynamic attention mechanism herein facilitates inter-modal calibration so that interpretability is guaranteed and false positives in low-quality signals are minimized. These benefits make ITG-Fusion a benchmark multimodal medical analytics model even with realistic hospital data limitations.

7. CONCLUSION AND FUTURE WORK

This paper introduced ITG-Fusion, a new multimodal deep learning architecture combining Vision Transformers (ViT), CNN-LSTM, and Temporal Graph Neural Networks (TGNN) using a Cross-Modal Attention Fusion (CMAF) mechanism to predict cardiovascular disease.

The model combines heterogeneous data sources—echocardiograms, ECG signals, and EHR records—into one, interpretable diagnostic pipeline that can learn and reason end-to-end.

Extensive experiments on three publicly available datasets (EchoNet-Dynamic, PTB-XL, and MIMIC-III) illustrate that ITG-Fusion attains state-of-the-art performance with an AUC of 0.982 and surpasses all the current baselines by 4–7 %.

The system also exhibits exceptional stability (Δ = 0.031) and cross-domain robustness with high predictive confidence, even with signal corruption and missing data.

Interpretability experiments (Figures 10–12) also establish that the fusion model makes clinically interpretable explanations, with more than 93 % agreement against cardiologist annotations.

The architectural organization ensures modular scalability, allowing real-time inference on edge devices like hospital servers or portable diagnostic units.

This flexibility places ITG-Fusion as an actionable foundation for tele cardiology that incorporates AI, filling the gap between multimodal deep learning research and clinical deployment.

Ethical Statement

This study utilizes publicly available and anonymized datasets (EchoNet-Dynamic, PTB-XL, and MIMIC-III). No direct human or animal subjects were involved. All procedures comply with ethical standards for secondary data usage and data privacy regulations.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REEFRENCES

[1] Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., & Zou, J. Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature, 580*(7802), 252–256. https://doi.org/10.1038/s41586-020-2145-8.

[2] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data, 7*, 154. https://doi.org/10.1038/s41597-020-0495-6.

[3] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035. https://doi.org/10.1038/sdata.2016.35.

[4] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine, 25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3.

[5] Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M., Andersson, C. R., Macfarlane, P. W., Meira, W., & Schön, T. B. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications, 11*, 1760. https://doi.org/10.1038/s41467-020-15432-4.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://papers.nips.cc/paper/6703.

[9] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[10] Alday, E. A. P., Gu, A., Shah, A., Robichaux, C., Wong, A. K. I., Liu, C., Liu, F., Rad, A. B., Elola, A., Seyedi, S., et al. (2021). Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement, 42*(9), 094001. https://doi.org/10.1088/1361-6579/ac1c4f.

[11] Hong, S., Yang, D., Fong, R., Kiyasseh, D., & Liu, F. (2022). Practical lessons on 12-lead ECG classification. *Frontiers in Physiology, 12*, 811661. https://doi.org/10.3389/fphys.2021.811661.

[12] Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., & Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *npj Digital Medicine, 6*, 74. https://doi.org/10.1038/s41746-023-00811-0.

[13] Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics, 23*(2), bbab569. https://doi.org/10.1093/bib/bbab569.

[14] Hayat, N., Khasanova, R., Erdem, A., Erdem, E., & Aubreville, M. (2022). MedFuse: Multimodal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the Machine Learning for Health (ML4H), PMLR 182* (pp. 212–229).

[15] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 787–795). https://doi.org/10.1145/3097983.3098126.

[16] Shang, J., Xiao, C., Ma, T., Li, H., & Sun, J. (2019). GAMENet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 1126–1133. https://doi.org/10.1609/aaai.v33i01.33011126.

[17] Lim, G. B. (2020). Estimating ejection fraction by video-based AI. *Nature Reviews Cardiology, 17*, 197. https://doi.org/10.1038/s41569-020-0375-y.

[18] PhysioNet. (2020). PTB-XL: A large publicly available electrocardiography dataset (resource page).

[19] Goldberger, A. L., et al. (2016). MIMIC-III Clinical Database (resource page). PhysioNet. https://physionet.org/content/mimiciii/1.4/

[20] He, B., et al. (2023). Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature, 616*, 520–524. https://doi.org/10.1038/s41586-023-05947-3.

[21] Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D.,

Bagul, A., Langlotz, C. P., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine, 15*(11), e1002686. https://doi.org/10.1371/journal.pmed.1002686 .

[22] Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: General overview. *Korean Journal of Radiology, 18*(4), 570–584. https://doi.org/10.3348/kjr.2017.18.4.570.

[23] Alday, E. A. P., et al. (2021). (Challenge dataset & outcomes) Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement, 42*(9), 094001. https://doi.org/10.1088/1361-6579/ac1c4f.

[24] Hong, S., et al. (2022). Practical lessons on 12-lead ECG classification. *Frontiers in Physiology, 12*, 811661. https://doi.org/10.3389/fphys.2021.811661.

[25] Wolf, T., et al. (2020). Transformers: State-of-the-art NLP (library paper). *EMNLP–Demos*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[26] Vaswani, A., et al. (2017). Attention is all you need (full paper PDF). In *NeurIPS 2017*. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[27] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs (full paper PDF). In *NeurIPS 2017*. https://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs.pdf.

[28] Taylor, C., & Nguyen, A. (2024). Heart failure and other cardiovascular diseases propel the demand for cardiac rhythm management markets. *Global Health Market Analytics*.

[29] Zwartkruis, V., Groenewegen, A., Rutten, F., Hollander, M., Hoes, A., & Van Der Ende, M. Y. (2020). Proactive screening for symptoms: A method to improve early detection of unrecognised CVD in primary care. *Preventive Medicine, 139*, 106143. https://doi.org/10.1016/j.ypmed.2020.106143

[30] Zhang, J., Yuan, M., & Wang, Y. (2022). Improving cardiovascular disease prediction using ensemble deep learning with clinical and imaging data. *Computer Methods and Programs in Biomedicine, 220*(1), 106815. https://doi.org/10.1016/j.cmpb.2022.106815

[31] Sargam, G. S., & Kalapala, R. (2025). A multi-modal federated graph learning approach for health insurance pricing with attention and explainability on the cloud. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291437

[32] Kalapala, R., & Sargam, G. S. (2025). Federated dual-modal anomaly detection on cloud for privacy-preserving health insurance fraud analytics. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291269

[33] Gorrepati, L. P., Kalapala, R., & Sargam, G. S. (2025). Leveraging artificial intelligence and big data in healthcare provider systems: Enhancing patient care and operational efficiency. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291497

[34] Kalapala, R., & Sargam, G. S. (2025). Personalized health insurance premium forecasting using AI: Behavioral and biometric data fusion with cloud computing on AWS for enhanced underwriting models. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291190

[35] Sargam, G. S., & Kalapala, R. (2025). AI-driven claim fraud detection in health insurance using federated anomaly detection networks with cloud computing on AWS for privacy-preserving financial security. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291290Jaric, S., & Petrova, A. (2023). Aptasensors in cardiovascular diagnostics: A

femtomolar approach. *Analytical Chemistry Innovations*.

[36] Ahmad, E., & Kumar, S. (2024). A review on cardiovascular biomarkers: Diagnostic and therapeutic perspectives. *Molecular Advances in Cardiology*.

[37] Omran, F., & Chatha, K. (2022). Cardiovascular biomarkers: Future directions and clinical applications. *International Journal of Cardiology and Molecular Sciences*.

[38] Ahmed, U., Lin, J. C.-W., & Srivastava, G. (2023). Towards early diagnosis and intervention: An ensemble voting model for precise vital sign prediction in respiratory disease. *IEEE Journal of Biomedical and Health Informatics*. Advance online publication. https://doi.org/10.1109/JBHI.2023.3270888

[39] Bai, J., Zhang, Z., Jin, W., Zhao, Y., Chen, H., & Zhou, S. (2025). A benchmark framework for the right atrium cavity segmentation from LGE-MRIs. *IEEE Transactions on Medical Imaging*. Advance online publication. https://doi.org/10.1109/TMI.2025.3590694

[40] Bianco, M., Scarciglia, A., Bonanno, C., & Valenza, G. (2025). Quantifying chaotic behavior in noisy dynamical systems: A study on heartbeat dynamics. *IEEE Transactions on Biomedical Engineering*. Advance online publication. https://doi.org/10.1109/TBME.2025.3566470

[41] Chen, C., Wang, Y., Zhang, J., Liu, X., Li, Y., & Zhang, S. (2025). High-quality CEST mapping with Lorentzian-model informed neural representation. *IEEE Transactions on Biomedical Engineering*. Advance online publication. https://doi.org/10.1109/TBME.2025.3574238

[42] Debus, L. Y., Kisse, J., Hergert, O., Rojek, K., Biebl, E. M., & Fischer, G. (2025). Blood makes a difference: Experimental evaluation of molecular communication in different fluids. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*. Advance online publication. https://doi.org/10.1109/TMBMC.2025.3602650

[43] Demirel, B. U., & Holz, C. (2025). Temporal cardiovascular dynamics for improved PPG-based heart rate estimation. *IEEE Journal of Biomedical and Health Informatics*. Advance online publication. https://doi.org/10.1109/JBHI.2025.3617604

[44] Eslami, N., & Moaiyeri, M. H. (2025). Efficient training and energy saving using ternary hardware acceleration for PCG classification. *IEEE Embedded Systems Letters*. Advance online publication. https://doi.org/10.1109/LES.2025.3547961

[45] Fang, J., Xu, Y., Wang, J., Zhou, X., Zhao, X., Zhang, H., … He, Q. (2025). Perspectives of global standards on validation of blood pressure measuring devices. *IEEE Transactions on Biomedical Engineering*. Advance online publication. https://doi.org/10.1109/TBME.2025.3587411

[46] Faes, L., Mijatovic, G., Sparacino, L., & Porta, A. (2025). Predictive information decomposition as a tool to quantify emergent dynamical behaviors in physiological networks. *IEEE Transactions on Biomedical Engineering*. Advance online publication. https://doi.org/10.1109/TBME.2025.3570937

[47] Fu, S., Wang, X., Li, Y., Chen, Z., Liu, Z., Yang, L., … Zhang, H. (2025). An ultrasound-guided real-time automatic navigation framework for magnetic guidewire robots to improve interventional surgery. *IEEE Transactions on Automation Science and Engineering*. Advance online publication. https://doi.org/10.1109/TASE.2025.3615271

[48] Kantu, N. T., Gao, W., Srinivasan, N., Buckner, G. D., & Su, H. (2025). Portable and versatile catheter robot for image-guided cardiovascular interventions. *IEEE/ASME Transactions on Mechatronics*. Advance online publication. https://doi.org/10.1109/TMECH.2025.3559911

[49] Li, G., Hu, W., & Lin, A. (2025). Modified Fourier-domain transfer entropy for cardiovascular analysis: Insights into aging mechanisms. *IEEE Journal of Biomedical and Health Informatics*. Advance online publication. https://doi.org/10.1109/JBHI.2025.3621184

[50] Liang, P., Du, Y., Qiao, Z., & Wang, S. (2025). CFTResNet: A novel cross-domain diagnosis framework guided by interpretability for cardiovascular diseases. *IEEE Journal of Biomedical and Health Informatics*. Advance online publication. https://doi.org/10.1109/JBHI.2025.3620820