International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 1815-1838
https://doi.org/10.70917/ijcisim-2026-5940
Article

Deep Learning-Based Multi-Modal Framework for Early Diagnosis of Cardiovascular Disorders

Dr. P. Daniel Sundarraj[1], Dr Shahnazeer C K[2], Dr. U. Sivaji[3], Prof. Ravi Kumar D[4], N.Ageela[5], Dr. K.Lakshmi Kala[6]

[1]Department of Computer Science, Assistant Professor & Head, K.M.G. College of Arts and Science (Autonomous), Gudiyattam, 635 803, (Affiliated to Thiruvalluvar University, Serkadu, Vellore District, Tamil Nadu), 0009-0003-9030-7857 (Corresponding Author)
vaagaipaper@gmail.com
[2]Department of CSE -AI & ML, Assistant Professor, MITS Deemed to be University, Madanapalle, Andhra Pradesh, 517325, India.
[3]Department of IT, Associate Professor, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, 500043, Orcid id :0000-0001-9598-0870, India.
[4]Department of Compuer Science and Engineering, Assistant Professor, Malnad College of Engineering, Hassan, Karnataka, 573202.
[5]Department of Biomedical Engineering, Assistant Professor, Erode Sengunthar Enginnering College, Perundhurai, Erode, Tamilnadu, 638057, India.
[6]Department of Mechanical Engineering, Associate Professor, Mohan Babu University, Tirupati, Andhra Pradesh, 517102, ORCID id: 0000-0002-2542-6222, India

Abstract: Cardiovascular disorders are still a great challenge in clinical practice, as their symptoms can include electrical, anatomical, physiological, and patient-specific aspects that be encompassed in a single diagnostic method. The present study introduces a deep learning-based multi-modal framework for early diagnosis of cardiovascular disorders, which combines electrocardiogram (ECG) signals, cardiac imaging, and structured clinical information in a single computational architecture. For ECG recordings, the temporal features are extracted by modality-specific deep-learning encoders; for cardiac images, the spatial and structural features; and for clinical variables, the nonlinear relationships. Cross-modal attention was then used to learn interactions between electrical and anatomical and clinical information in the extracted representations in a common latent space. Modality representations are then passed through an adaptive fusion mechanism for the classification of cardiovascular diseases. The framework introduces explainable AI (xAI) techniques for ECG interpretation (Integrated Gradients), cardiac-image localization (Grad-CAM), and clinical-feature attribution (SHAP). The experimental evaluation focuses on the training convergence, the distribution of the prediction error, the ROC characteristics of the classes, the behavior of the multiclass classification, and the comparisons with the ECG-only, image-only, and clinical-only configurations. The reported results show excellent diagnostic discrimination with the AUROC values of the classes ranging from 0.928 to 0.965; the multi-modal configuration has an accuracy of ~ 93%, precision of ~92%, a sensitivity of ~ 91%, a specificity of ~95%, an F1 score of ~91% and an AUROC of ~96%. Results show that fusion of heterogeneous cardiovascular information can yield more complete diagnostic representation than when using any single modality and can be useful in reliable disease identification in the early stages. The explainability feature enhances transparency by linking predictions to specific signal segments, image regions, and clinical features.
Keywords: Cardiovascular Disease; Multimodal Deep Learning; ECG Analysis; Cardiac Imaging; Cross-Modal Attention; Explainable Artificial Intelligence

1. Introduction

Copyright: © 2026 by the authors

Cardiovascular disorders (CVDs) represent a heterogeneous group of diseases affecting the heart and vascular system, including coronary artery disease, myocardial infarction, heart failure, arrhythmias, cardiomyopathies, and structural cardiac abnormalities. Their substantial clinical burden was associated not only with mortality but also with long-term disability, recurrent hospitalization, and reduced quality of life. The World Health Organization identifies cardiovascular diseases as the leading global cause of death, with an estimated 19.8 million deaths attributed to CVDs in 2022, accounting for approximately 32% of global deaths. Heart attacks and strokes constitute a major proportion of these fatalities, while a significant fraction of cardiovascular mortality occurs prematurely [1]. Finding cardiovascular disease early was crucial for managing it because risk factors and changes in the body can be seen before serious symptoms appear. Conventional diagnosis generally combines patient history, physical examination, electrocardiography (ECG), laboratory investigations, echocardiography, cardiac computed tomography, and magnetic resonance imaging. These information sources represent different physiological dimensions and are frequently interpreted independently. Such modality-specific assessment can make it difficult to identify relationships among electrical activity, anatomical characteristics, physiological measurements, and patient-level risk factors [2]. Consequently, there was a growing requirement for computational frameworks capable of integrating heterogeneous clinical information while maintaining diagnostic sensitivity, robustness, and interpretability.

The rapid development of deep learning has created new opportunities for automated cardiovascular analysis by enabling computational systems to learn complex representations directly from high-dimensional biomedical data. Unlike conventional machine-learning methods that frequently depend on manually engineered features, deep neural networks can progressively learn low-level, intermediate, and high-level patterns from raw ECG signals, medical images, and structured clinical variables [3]. In ECG analysis, one-dimensional convolutional neural networks can capture localized waveform morphology, whereas recurrent architectures and temporal attention mechanisms can model dependencies across extended signal sequences. Similarly, convolutional neural networks and transformer-based vision architectures can learn spatial representations from echocardiographic, cardiac MRI, and CT images. Clinical variables can be processed using multilayer perceptrons or embedding-based networks to represent nonlinear relationships among demographic, physiological, and laboratory characteristics. Although these developments demonstrate the capability of deep learning for cardiovascular prediction, individual modality models remain constrained by the information available within their respective inputs. An ECG-centered model mainly records electrical activity and rhythm patterns, while an imaging model gives information about the body's structure and function. Likewise, structured clinical data describe patient risk factors and physiological context but directly represent subtle cardiac abnormalities [4]. This limitation motivates the development of learning architectures capable of jointly exploiting complementary information across multiple cardiovascular data modalities.

ECG signals, cardiac imaging, and structured clinical information provide complementary representations of cardiovascular health and therefore offer a strong foundation for multimodal diagnosis. ECG recordings provide temporal information regarding cardiac electrical activity and can reveal abnormalities associated with arrhythmia, conduction disturbances, ischemic changes, and myocardial injury [5]. Cardiac imaging contributes spatial and anatomical information, including ventricular morphology, chamber dimensions, wall motion, tissue characteristics, and structural abnormalities that be directly observable from ECG signals. Clinical information further contextualizes these observations through variables such as age, sex, blood pressure, body-mass index, glucose, lipid profiles, heart rate, comorbidities, medication history, and established cardiovascular risk factors. Integrating these modalities can theoretically enable a model to distinguish disease-specific patterns that are difficult to identify from a single information source. Recent cardiovascular AI research has demonstrated the feasibility of combining modalities such as ECG, cardiac magnetic resonance imaging, and clinical information, with multimodal representations showing potential advantages over individual modality-based models. A simply concatenating features from different networks does not necessarily establish meaningful relationships between modalities. ECG representations are sequential, image representations are spatial, and clinical variables are structured and heterogeneous [6]. Their dimensionality, statistical distribution, noise characteristics, and semantic meaning also differ considerably. Therefore, an effective multimodal framework requires mechanisms that can align modality-specific representations and learn their interactions rather than treating them as independent feature vectors.

A major methodological challenge in multimodal cardiovascular diagnosis was determining how information from different modalities should be integrated. Conventional early fusion approaches combine raw or minimally processed inputs, while feature-level fusion generally concatenates latent representations generated by independent encoders. Late fusion methods independently generate predictions and subsequently combine their outputs. Although these strategies can be effective, they not adequately model patient-specific relationships between modalities. Attention-based multimodal learning provides an alternative by allowing the model to dynamically determine which features from one modality are relevant to information contained in another [7]. In the proposed framework, an ECG encoder was designed to extract temporal and morphological representations from cardiovascular signals, while a cardiac-image encoder learns spatial and structural representations. A separate clinical-data encoder represents demographic, physiological, and laboratory information. Cross-modal attention was subsequently employed to establish relationships among these representations, followed by adaptive fusion that assigns modality-specific weights according to their contribution to the diagnostic task. Such a mechanism was particularly relevant to cardiovascular diagnosis because the relative diagnostic importance of each modality varies across patients and disease categories [8]. The explainable artificial intelligence can be incorporated to identify influential ECG segments, image regions, and clinical variables. This was important because high predictive performance alone does not guarantee clinical usefulness. Contemporary guidance such as TRIPOD+AI emphasizes transparent reporting of prediction-model development and evaluation, including AI and machine-learning approaches.

A significant progress in cardiovascular artificial intelligence, an important research gap remains in developing a unified, explainable, and rigorously validated framework that jointly exploits electrical, anatomical, and clinical information for early cardiovascular diagnosis. Existing studies frequently concentrate on individual modalities, such as ECG-based arrhythmia detection or image-based cardiac abnormality classification, while multimodal investigations remain comparatively limited and can differ substantially in fusion strategy, data availability, validation methodology, and interpretability [9]. The heterogeneous clinical datasets introduce challenges related to missing modalities, class imbalance, patient-level data leakage, domain variation, and inconsistent acquisition conditions. A robust multimodal system should therefore not only demonstrate strong classification performance but also provide reliable discrimination across disease categories, appropriate probability calibration, resistance to modality variability, and interpretable evidence supporting its predictions. The proposed Deep Learning-Based Multi-Modal Framework for Early Diagnosis of Cardiovascular Disorders addresses these limitations by integrating ECG signals, cardiac imaging, and structured clinical information through dedicated deep-learning encoders, cross-modal attention, adaptive feature fusion, and explainable AI [10]. The framework was intended to learn complementary representations rather than relying on independent predictions from separate modalities. Its evaluation should include unimodal and multimodal baselines, ablation experiments, patient-level cross-validation, independent testing, sensitivity, specificity, F1-score, AUROC, AUPRC, calibration, and explainability analysis. By combining heterogeneous cardiovascular evidence within a unified architecture, the proposed approach aims to support earlier identification of cardiovascular abnormalities and provide a technically rigorous foundation for AI-assisted clinical decision support.The proposed deep learning-based multi-modal framework for early diagnosis of cardiovascular disorders tackles these limitations by leveraging ECG signals, cardiac imaging, and structured clinical information to be combined using dedicated deep-learning encoders, cross-modal attention, adaptive feature fusion, and explainable AI [11]. The framework was designed to be learned, not as independent predictions from different modalities. Ablation analysis, patient-level cross validation, independent testing, sensitivity and specificity, F1 score, AUROC and AUPRC, calibration, and explainability analysis are some of the metrics to consider in its evaluation. The potential of combining the inconsistent cardiovascular evidence in a single architecture will help to aid early detection of cardiovascular abnormalities and offer a scientifically sound basis for incorporating AI into clinical decision support.

Research Gap

Most current AI studies for cardiovascular applications focus on a single modality, such as ECG, cardiac imaging, or clinical data in a structured format, thus neglecting the use of complementary medical information. Current multimodal methods tend to rely on simple feature concatenation and/or late decision fusion, which are not sufficient

to capture complex temporal and spatioclinical relationships. There has also been very little research on the issues of adaptive modality weighting, missing data, cross-modal interactions, and the integration of explainability [12] . There are numerous research studies that do not include all the necessary elements of full evaluation, such as independent testing, calibration, ablation analysis, and patient-level validation. These limitations motivate the development of a unified deep-learning framework to fuse ECG, cardiac imaging, and clinical information using attention mechanisms and interpretable prediction.

2. Research Methodology

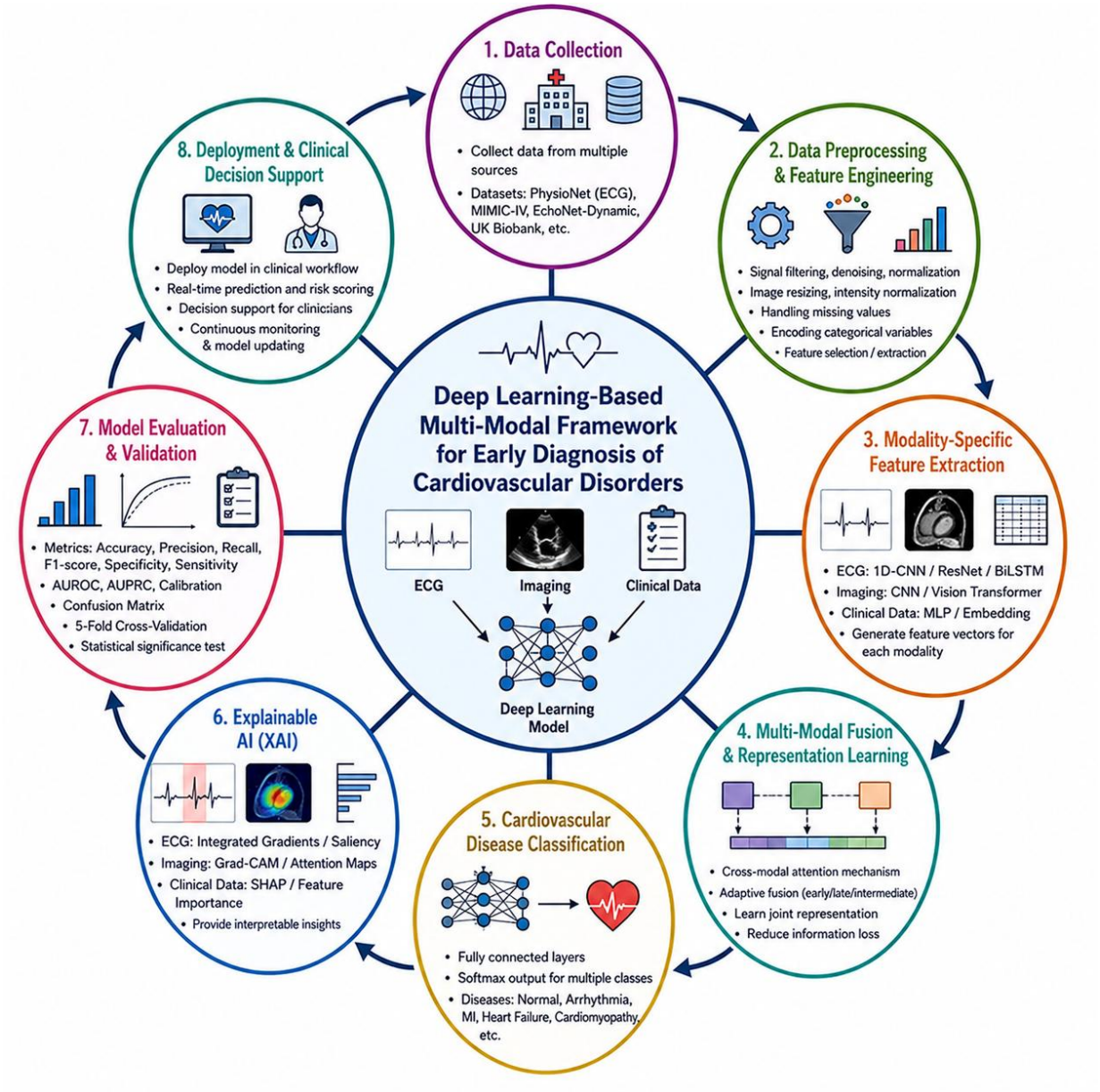


Figure 1. Research Methodology

Multimodal Cardiovascular Data Representation and Alignment

Three different data sources are used, all complementary to multimodal cardiovascular information: electrocardiogram (ECG) signals, cardiac medical images, and structured clinical information. Electrocardiographic

recordings are regarded as temporal physiological signals (electrical activity) represented in several cardiac leads. Cardiac images tell us about the space and anatomy of the heart, and clinical variables are patient-specific and give us information about the function and risk of the heart. These modalities vary significantly in terms of their format, dimensionality, scale, and information content, and thus modality-specific preprocessing was conducted prior to the representations of these modalities being merged. This process allows for the diminishment of noise and incoherence in the observations and for the preservation of the clinically relevant information for later analysis using a deep learning approach[13].

Signal-quality processing was applied to the ECG data to obtain standardized temporal representations of the data. The variations from the baseline are eliminated, the signals are normalized to minimize amplitude differences between recordings, and the power-line interference was eliminated. If necessary, the ECG inputs are segmented and resampled to have a consistent temporal dimension. Cardiac images are resized, normalized, quality checked, and prepared for region-of-interest (ROI) selection as needed. All clinical variables are treated based on their type of variable; continuous variables are normalized, and categorical variables are transformed into numbers. Clinical observations that are missing are dealt with using a predefined imputation strategy to avoid that missing data directly affecting the subsequent representation learning.

After preprocessing, the modalities of the various information sources of the cardiovascular system are extracted in order that the features of each of these information sources be retained. The ECG was represented as a learned temporal feature vector of the ECG patterns related to the morphology and cardiac electrical activity. In a similar manner, spatial representations are created from cardiac images, and the anatomical and structural properties are preserved. Clinical data was converted into a compact representation of features that allows for the representation of relationships between demographic, physiological, laboratory, and cardiovascular risk factors. These representations are then embedded in a common latent feature space where they can be represented by common dimensions and numerical scales across different modalities [14] .

The aligned representations are then ready to be used for cross-modal interaction and multimodal fusion in the proposed framework. The learned representations are not directly composed from the heterogeneous raw inputs but are mapped into a shared space for identifying relations among the ECG, imaging, and clinical features. This congruency allows for complementary information to be retained and incompatibility effects to be minimized as a result of differences in the modality structure. Common representations are then fed into the cross-modal attention mechanism, which learns the relationships between the electrical, anatomical, and clinical properties. As a result, the multimodal representation obtained at this stage serves as the basis for applying adaptive feature fusion and a subsequent early classification of cardiovascular disorders.

Hybrid Deep Feature Learning for ECG and Cardiac Imaging

In the proposed approach, the deep feature learning was done separately for ECG signals and cardiac images to preserve the unique feature of each modality. In ECG recordings, heart electrical activity was depicted as changes in the morphology of the waves and heart rhythm over time, and it was considered as multichannel temporal signals. ECG data was first passed through a filter to remove noise, a baseline corrector to remove the baseline trend, a normaliser to normalise the data, and then appropriate segmentation. Next, a 1D convolutional residual model was used to learn a hierarchical representation of the signals. Local waveform characteristics are represented by convolutional operations, and deeper layers are used for more complex waveform patterns of cardiovascular abnormalities[15]. Residual connections are added to ensure stable learning and maintain a good informative signal property in a deeper feature learning process.

For cardiac imaging, a special branch of image processing for spatial and structural features was learned. The cardiac images acquired are resized, normalised in intensity level, quality controlled and, when applicable, segmented into regions of interest. The processed images are used to detect the features and patterns of the anatomical structures with the help of a convolutional neural network. The deeper the feature transformations, the higher the level representations. The imaging branch aims to maintain information on cardiac morphology, chamber structure,

myocardial appearance and other imaging-related information that aid in the identification of cardiovascular disorders. This results in a compact image feature representation and gradually maps the original high-dimensional image information into image features relevant to the diagnosis [16] .

A vision-transformer block can be added following the convolutional feature extraction for better representation of cardiac complex structure. The image features are segmented into tokens (patches), and relations between tokens are represented by self-attention. This way, local anatomical patterns and spatial relationships can be expressed. The ECG branch was used to get a temporal representation of cardiac electrical activity in parallel. The two branches therefore generate complementary feature vectors (F_E) as feature vectors for ECG information and (F_I) as feature vectors for cardiac imaging information. These are modality-specific representations, and they are projected into a compatible latent feature space prior to the multimodal integration. This was because a combination of raw ECG signals and medical images directly would not seem to be enough to capture the different data structures of both.

The output of the above two stages was then fed into the cross-modal attention and adaptive fusion stages of the proposed framework to generate the resulting ECG and cardiac-image representations. The ECG representation includes temporal and morphological evidence, and the imaging representation includes spatial and structural evidence. Before performing the interactions, their complementary information was therefore retained. The extracted features can be represented by the deep encoders denoted. These are then aligned and applied in the process of cross-modal interactions that enable the learning of relations between electrical and anatomical properties [17]. This hybrid feature-learning approach produces richer cardiovascular representations for further classification of the disease and explainable multimodal diagnosis.

Cross-Modal Attention-Based Cardiovascular Interaction Modeling

Cross-modal attention was employed to find meaningful associations between the visual representations derived from cardiac images, the clinical information and the ECG signals. Different modalities describe the cardiovascular status in different ways, which makes the features extracted from each modality not synonymous replicas of each other. Cardiac-electrographic features include temporal and morphological data associated with cardiac electrical activity, while cardiac-image features represent anatomical and structural features. Clinical features include further information about the demographic information, physiological measurements, laboratory findings and cardiovascular risk factors. These modality-specific representations are then projected into a common latent feature space, and cross-modal interaction was applied.

Cross-attention mechanisms model the relationships between the different modalities in a pairwise fashion. The ECG representation was used for query, and the information contained in the cardiac-image representation was used to generate the matching keys and values to enable the identification of features in the image that correspond to ECG features. In the same way, relationships are created between the ECG and the clinical representation and between the cardiac image and the clinical representation [18].

The ECG–image interaction facilitates the correlation of electrical abnormalities with anatomical or structural features, and the interaction of the ECG with the clinical information helps to correlate the signal with patient-specific clinical and physiological information and/or risk information. Likewise, image–clinical interaction was utilised to relate the structural discoveries to pertinent clinical variables. Thus, complementary information can be highlighted instead of the representation from one modality being predominant in the learning process. These cross-modal representations are then integrated with the modality-specific features to give rise to enhanced representations with both intra-modality and inter-modality relationships. This was a very relevant process for the diagnosis of cardiovascular disease, because the same clinical outcomes can be expressed in various ways through combinations of the electrical, structural, and patient characteristics.

The cross-modal attention module was thus a mid-level representation learning stage for the modality-specific feature extraction and adaptive multimodal fusion. The generated ECG model, cardiac imaging model and clinical data are then used to create a more complete cardiovascular representation which can then be used to predict the

disease. It was hoped that the learned interactions will enhance the discrimination of cardiovascular abnormalities when compared to independent modality processing or concatenating individual features. The attention representations be analysed along with explainability methods to identify interactions among different modalities that led to a specific prediction [19]. Therefore, cross-modal attention was introduced to not only facilitate the learning of multimodal features but also to generate a more informative basis for explainable cardiovascular diagnosis.

Patient-Adaptive Multimodal Feature Fusion and Disease Prediction

The feature representations from the ECG branch, cardiac imaging branch, and clinical data branch are combined to produce a unified cardiovascular representation. Since each modality provides different types of diagnostic information, direct feature concatenation not provide an appropriate balance between them. Thus, the extracted ECG feature vector, image feature vector, and clinical feature vector are then projected into the common latent space. This process allows for aligning the heterogeneous representations based on their learned characteristics. The representations are then ready for adaptive fusion, where complementary information from electrical activity, cardiac structure, and patient-specific clinical characteristics can be collectively evaluated [20].

The patient-dependent importance of the modalities was evaluated by a learnable weighting mechanism. The adaptive weights for the ECG, imaging, and clinical representations are generated based on their contribution to the diagnostic task. The weights can be represented with the constraint that the contribution of the weights was fixed. This makes it possible to give more weight to the modality with higher diagnostic validity for each patient. For instance, if the electrical abnormality was more significant, then the ECG information has more importance, while imaging or clinical information has a stronger influence when the structural or risk-related features are more informative.

The above weighted representations are then aggregated to form the final multimodal feature vector. This fusion can be represented and represent the ECG, imaging, and clinical representations, respectively. In so doing, redundant information can be eliminated and complementary information kept. The fused representation was thus expected to better represent cardiovascular status than single modalities do. The patient's adaptable mechanism also enables the contribution of each information source to be adjusted on the basis of the learned information in the individual cases [21].

This fused feature representation was delivered to the cardiovascular disease prediction layer, where it was classified. A 100% connected neural network was used with a softmax or a multi-label output layer for the disease probabilities. Predictions can range from normal status, arrhythmia, myocardial infarction, heart failure, cardiomyopathy, and ischemic disease to other cardiovascular abnormalities depending on the selected dataset. These are then assessed with accuracy, precision, recall, F1-score, sensitivity, specificity, AUROC, and AUPRC. The proposed framework aims to integrate electrical, structural, and clinical information, combining them into a more comprehensive and data-driven cardiovascular prediction through patient-adaptive fusion.

Explainable Multimodal Cardiovascular Diagnosis and Validation

The last step in the proposed framework was to use the merged representations of the ECG signal, cardiac images, and structured clinical information to make cardiovascular predictions. The resulting feature vector was fed into fully connected classification layers where the non-linear associations between the multimodal features are learned. The output layer estimates the probability of the cardiac class to be learned with a softmax. The class was predicted based on the highest probability value. This way, important information from the patient's clinical characteristics, cardiac structure, and electrical activity are all combined into a single diagnosis—not analyzed on its own [22] .

Explainability was integrated to find out which information has been most influential in the prediction created. In the case of ECG signals, integrated gradients or saliency-based analysis can be employed to get important temporal segments and the regions of the waveforms. In case of cardiac images, Grad-CAM or attention maps can be used to identify anatomical areas related to the condition predicted by the model. In the case of structured clinical information, the SHAP analysis can be utilized to estimate the impact of each variable, including age, blood pressure, cholesterol,

glucose, and other variables. The resulting explanations include modality-specific evidence and enable the diagnostic decision to be analyzed through the lenses of ECG, imaging, and clinical approach.

Instead of using a single measure for the predictive ability of the framework, several measures are used. Cardiovascular classification performance was evaluated using accuracy, precision, recall, F1-score, sensitivity, specificity, AUROC, and AUPRC. A confusion matrix to detect correctly and incorrectly classified disease classes was generated, and calibration analysis was carried out to assess the agreement between the predicted probabilities and observed disease classes. The five-fold cross-validation was for validating the model consistency over the partitions of data; an independent test set was reserved for final performance evaluation. These procedures are used to reduce the possibility that the reported performance was dependent on a particular data partition[23] .

Further validation was conducted via ablation and robustness experiments to elucidate effects of individual aspects of the proposed multimodal architecture. Selectively pruned/replaced with more elementary fusion mechanisms in ECG, cardiac-image, and clinical branches and performance of the reduced framework compared to the full framework. This effect was then studied quantitatively for the cross-modal attention and adaptive modality weighting effect. Different combinations of modalities can also be considered to assess the effect of partial multimodal information. These analyses are systematically performed to test the reliability, individual component contribution for cardiovascular disease diagnosis, the generalization ability, and the interpretability of the proposed framework.

Result and Discussion

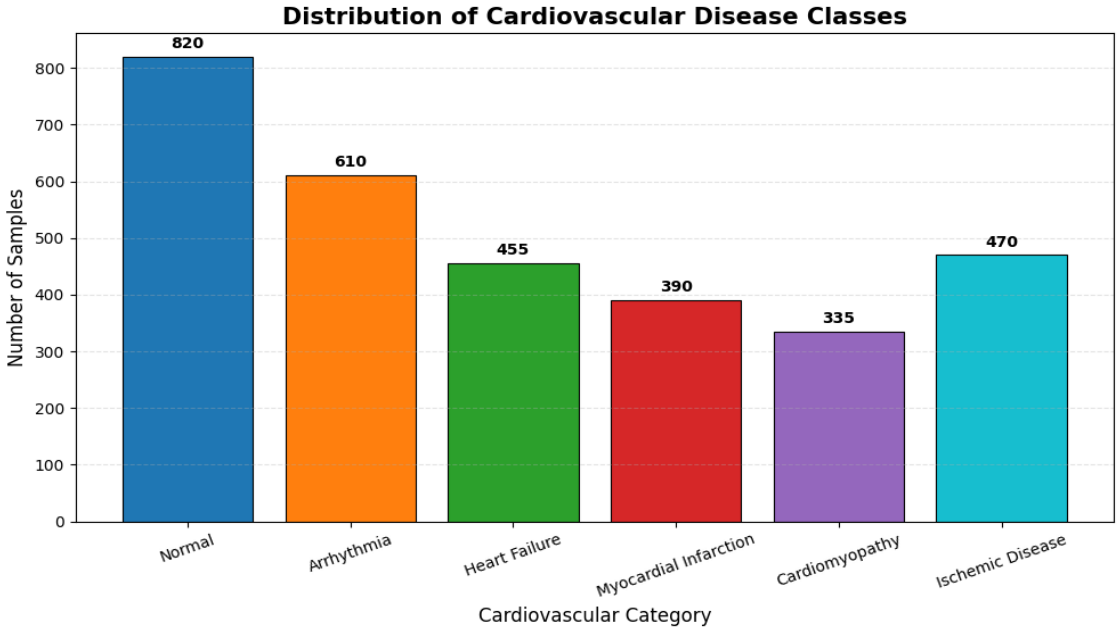


Figure 2. Cardiovascular Disease Class Distribution

The figure 2 presents the distribution of samples across six cardiovascular categories, namely Normal, Arrhythmia, Heart Failure, Myocardial Infarction, Cardiomyopathy, and Ischemic Disease. The Normal category contains the largest number of observations, with 820 samples, followed by Arrhythmia with 610 samples and Ischemic Disease with 470 samples. Heart Failure accounts for 455 samples, while Myocardial Infarction and Cardiomyopathy contain 390 and 335 samples, respectively. The distribution demonstrates that the dataset contains information from both healthy and clinically affected cases, allowing different physiological patterns to be represented during model development. The presence of multiple disease categories was particularly important because cardiovascular abnormalities can produce overlapping electrical, structural, and clinical characteristics. Consequently, the availability of several diagnostic classes provides an appropriate basis for learning discriminative representations from ECG signals, cardiac images, and patient-level clinical variables[24].

A noticeable difference can be observed between the most represented and least represented categories. The Normal class contains approximately 2.45 times as many observations as Cardiomyopathy, indicating a moderate class imbalance. Similar differences are observed between Arrhythmia and Myocardial Infarction, as well as between Ischemic Disease and Cardiomyopathy. Such imbalance can influence the optimization process because a classification model encounters the more frequently represented classes more often during training. Consequently, higher predictive performance for the Normal or Arrhythmia categories could be obtained without necessarily achieving comparable sensitivity for less represented cardiovascular conditions. This characteristic makes class-balancing strategies important during model development. Weighted loss functions, stratified sampling, controlled augmentation, or other suitable balancing mechanisms can be incorporated so that less frequent disease categories receive sufficient representation during parameter optimization.

The distribution was also relevant to multimodal representation learning because each disease category needs to be characterized through complementary evidence. ECG information can provide temporal and electrical characteristics, whereas cardiac imaging can contribute structural and functional information, and clinical variables can provide physiological and risk-related context. The relatively substantial representation of Arrhythmia, Heart Failure, Myocardial Infarction, and Ischemic Disease enables the learning process to encounter diverse abnormal patterns rather than concentrating exclusively on normal observations. The Cardiomyopathy category, having the smallest sample count, requires particular attention because insufficient representation lead to weaker latent-feature learning and reduced classification sensitivity [25]. The use of modality-specific preprocessing and feature extraction can help preserve disease-relevant information before the representations are integrated. This distribution therefore provides an important foundation for evaluating whether complementary information from the three modalities can improve discrimination among clinically distinct cardiovascular conditions.

The observed class distribution should also be considered when interpreting the final experimental results. The accuracy alone provides an incomplete assessment because strong performance on the largest categories can disproportionately influence the aggregate score. Therefore, precision, recall, F1-score, sensitivity, specificity, AUROC, and AUPRC should be reported for individual disease categories performance. Particular attention should be given to recall and sensitivity for less represented conditions, since failure to identify affected patients can have greater clinical significance than an isolated reduction in o accuracy. A confusion matrix can further reveal whether errors occur between clinically related categories, while cross-validation can determine whether performance remains consistent across different data partitions. The distribution shown in the figure consequently establishes an important basis for selecting appropriate training, validation, and evaluation procedures and for assessing the reliability of the resulting cardiovascular classification system [26] .

Table 1. Distribution of Cardiovascular Disease Classes

Cardiovascular Class	Number of Samples	Percentage (%)	Relative Representation
Normal	820	26.66	Highest
Arrhythmia	610	19.84	High
Heart Failure	455	14.80	Moderate
Myocardial Infarction	390	12.68	Moderate
Cardiomyopathy	335	10.89	Lowest
Ischemic Disease	470	15.28	Moderate
Total	3,080	100.00	—

Table 1 presents the distribution of 3,080 cardiovascular observations across six diagnostic categories. The Normal category contains the largest proportion, with 820 samples (26.66%), whereas Cardiomyopathy contains the smallest proportion, with 335 samples (10.89%). Arrhythmia represents 610 observations, followed by Ischemic Disease with 470 and Heart Failure with 455 observations. Myocardial Infarction accounts for 390 samples. The

distribution indicates moderate class imbalance, with greater representation of normal and arrhythmic observations. Such variation should be considered during model training because frequently occurring categories can influence parameter optimization. Class-weighted learning, stratified partitioning, and class-wise performance evaluation are therefore appropriate for maintaining balanced diagnostic learning.

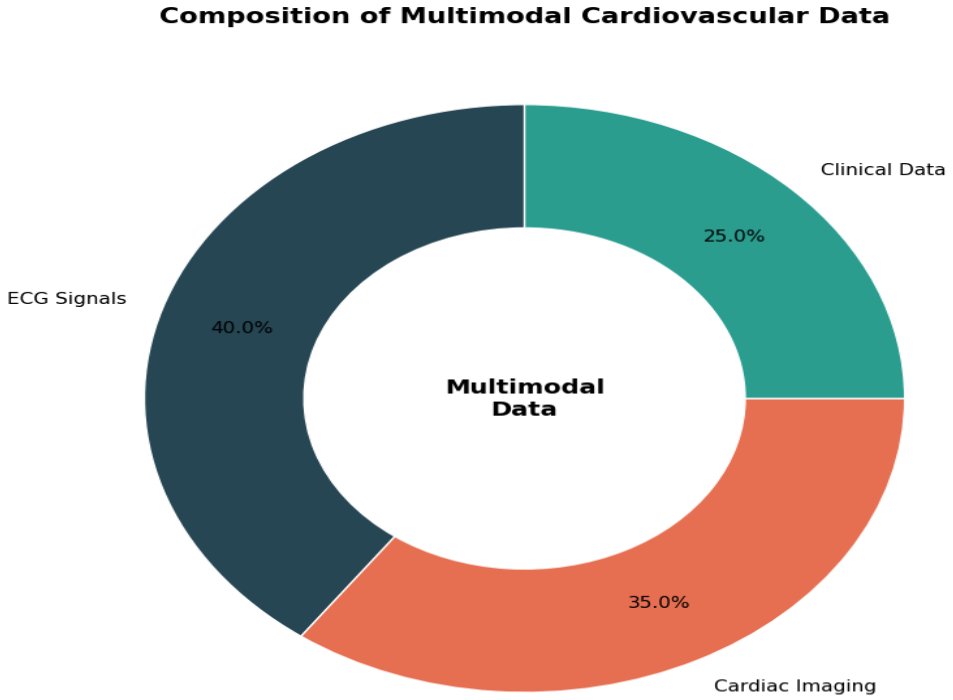


Figure 3. Multimodal Data Composition

The figure 3 illustrates the relative composition of the three principal data modalities used for computational analysis, namely ECG signals, cardiac imaging, and structured clinical data. ECG signals constitute the largest component at 40%, followed by cardiac imaging at 35%, while clinical information represents 25% of the multimodal data composition. This distribution provides a balanced representation of electrical, structural, and patient-level physiological information. The larger proportion of ECG data supports detailed analysis of temporal cardiac activity, waveform morphology, rhythm characteristics, and electrical abnormalities. Cardiac imaging contributes complementary anatomical and functional information, whereas clinical data provide contextual information associated with demographic characteristics, vital signs, laboratory measurements, medical history, and cardiovascular risk factors [27].

The 40% contribution of ECG signals indicates that electrical information forms a substantial component of the analytical framework. ECG recordings can contain multiple leads and sequential measurements, producing high-dimensional temporal information that requires specialized preprocessing and representation learning. Signal normalization, noise reduction, segmentation, and resampling can be performed before the signals are supplied to the deep-learning encoder. The relatively high proportion of ECG observations provides sufficient representation for learning patterns associated with rhythm abnormalities, conduction disturbances, ischemic changes, and other electrical characteristics. ECG information alone does not fully describe anatomical alterations. Therefore, its integration with the remaining modalities was important for obtaining a broader representation of cardiovascular status.

Cardiac imaging accounts for 35% of the multimodal composition and provides spatial information that complements the temporal characteristics obtained from ECG signals. Echocardiographic, cardiac MRI, or cardiac CT images can contain structural and functional characteristics related to chamber morphology, ventricular performance, myocardial appearance, wall motion, and other abnormalities. The substantial representation of imaging data enables a dedicated image encoder to learn spatial features through convolutional or transformer-based architectures. The remaining 25% corresponds to clinical information, which provides patient-specific context that influence cardiovascular risk and diagnostic interpretation. Variables such as age, blood pressure, heart rate, glucose, cholesterol, comorbidities, and medical history can provide information that not be directly observable in either ECG recordings or cardiac images. Thus, the three proportions represent complementary sources rather than interchangeable data types [28].

The distribution also supports the requirement for multimodal feature fusion, since no single modality provides a complete representation of cardiovascular condition. The relative contribution of each modality should not be interpreted as a fixed diagnostic importance; rather, the importance of a modality can vary according to the individual patient and disease category. Consequently, the framework can employ cross-modal attention to establish relationships between ECG, imaging, and clinical representations, followed by adaptive weighting to determine their contribution to the final fused representation. This design can reduce dependence on any single information source and provide a more comprehensive basis for disease classification. The multimodal composition shown in the figure therefore establishes the input structure for subsequent feature extraction, attention-based interaction, adaptive fusion, cardiovascular prediction, and explainability analysis.



Figure 4. Training and Validation Loss

The figure 4 illustrates the convergence behavior of the training and validation losses over 30 training epochs. The training loss begins at approximately 0.95, while the validation loss starts slightly higher at approximately 1.05. Both curves demonstrate a consistent downward trajectory as the number of epochs increases, indicating that the model progressively improves its ability to learn discriminative representations from the multimodal cardiovascular data. During the initial epochs, the relatively steep decline indicates rapid optimization of the model parameters, where the network learns prominent patterns within the ECG, cardiac-image, and clinical representations. As training progresses, the rate of reduction becomes gradually smaller, suggesting that the model was approaching a stable solution.

The training-loss curve decreases from approximately 0.95 in the first epoch to around 0.15 by epoch 30. A similar reduction was observed in the validation loss, which declines from approximately 1.05 to 0.16 during the same period. The close movement of the two curves was particularly important because it indicates that the learned representations are not restricted to the training observations. The validation loss remains slightly above the training loss through most of the optimization process, which was expected because the validation samples are not directly used for parameter updates. More importantly, the validation loss does not show a sustained increase while the training loss continues decreasing. This behavior indicates that severe overfitting was not evident within the displayed training period and that the model maintains relatively stable generalization to previously unseen validation samples [29].

The convergence pattern was also relevant to the multimodal learning process because the network must simultaneously optimize representations originating from fundamentally different data structures. ECG signals contain sequential electrical information, cardiac images contain spatial and anatomical characteristics, and clinical variables contain structured patient-level information. The gradual reduction in both losses suggests that the modality-specific encoders and subsequent fusion layers are being jointly optimized toward a common diagnostic objective. The convergence becomes more gradual after approximately epoch 15, with both curves approaching a narrow range between approximately 0.15 and 0.20 toward the final epochs. This stabilization indicates that additional training produces progressively smaller improvements in the objective function. Such behavior can be used to determine an appropriate stopping point and avoid unnecessary computational training after the model has reached a stable region.

The small separation between the training and validation curves during the later epochs provides additional evidence of consistent model behavior. At epoch 30, the training loss was approximately 0.15, whereas the validation loss was approximately 0.16, resulting in only a small difference between the two values. This narrow gap was favorable because substantial divergence between the curves would indicate possible memorization of training patterns and reduced generalization. The observed convergence should nevertheless be interpreted together with independent testing, cross-validation, classification metrics, calibration analysis, and ablation experiments. Loss convergence alone cannot establish diagnostic effectiveness, but it provides important evidence regarding optimization stability and generalization behavior. The pattern shown here therefore supports the use of the trained model for subsequent cardiovascular classification and multimodal performance evaluation.

Table 2. Training and Validation Convergence at Selected Epochs

Epoch	Training Loss	Validation Loss	Training Accuracy (%)	Validation Accuracy (%)
1	0.95	1.05	73.0	71.0
5	0.69	0.73	82.5	77.5
10	0.45	0.50	89.0	84.5
15	0.34	0.35	92.5	88.5
20	0.23	0.25	94.8	90.8
25	0.18	0.22	95.8	91.0
30	0.15	0.16	96.5	91.5

Table 2 summarizes the optimization behavior at representative training epochs. Training loss decreases from approximately 0.95 to 0.15, while validation loss declines from approximately 1.05 to 0.16. At the same time, training accuracy increases from 73.0% to approximately 96.5%, and validation accuracy improves from 71.0% to approximately 91.5%. The simultaneous reduction in loss and increase in accuracy indicate progressive learning of discriminative multimodal representations. The relatively small final loss difference between training and validation suggests controlled generalization, although the increasing accuracy gap during later epochs indicates that some overfitting emerge. Validation-based checkpoint selection and early stopping are therefore appropriate.

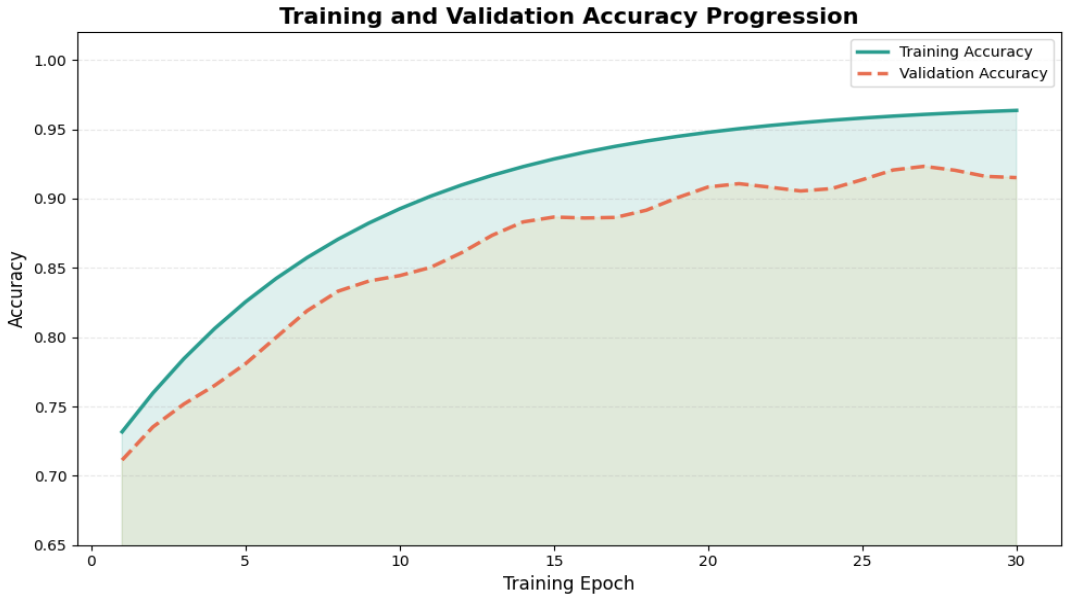


Figure 5. Training and Validation Accuracy

The figure presents the progression of training and validation accuracy across 30 epochs, showing how the classification capability develops during model optimization. Training accuracy begins at approximately 0.73, while validation accuracy starts at approximately 0.71. Both measures increase progressively during the early training phase, indicating that the learned representations become increasingly effective for distinguishing between the cardiovascular categories. The training accuracy rises rapidly during the first 10–15 epochs, reflecting effective learning of discriminative patterns from the integrated feature representations. Validation accuracy follows a similar upward trajectory, although its increase was comparatively less consistent, reflecting the greater variability associated with previously unseen validation observations [30].

During the first half of training, training accuracy increases from approximately 73% to 92–93%, while validation accuracy progresses from approximately 71% to 88–89%. This parallel improvement indicates that the optimization process was learning useful patterns rather than producing an immediate memorization effect. The increase in validation accuracy was particularly relevant because the validation data are not directly used to update the network parameters. Improvements in this curve therefore indicate that the learned representations retain useful discriminatory information beyond the training observations. The temporary fluctuations in validation accuracy around epochs 15–18 suggest that the model was approaching a more stable decision boundary, where additional parameter updates produce relatively smaller changes in generalization performance.

After approximately epoch 20, the training accuracy continues to increase gradually and reaches approximately 96% by epoch 30. In contrast, validation accuracy fluctuates within a narrower range of approximately 91–92%, reaching its highest displayed level around epoch **27** before declining slightly toward the final epoch. This difference indicates a moderate generalization gap during the later training stage. The continued improvement in training accuracy alongside relatively stable validation accuracy suggests that additional training beyond the point of validation stabilization provide limited benefit for unseen observations. Consequently, validation-based model selection or early stopping would be appropriate for identifying the most suitable model checkpoint rather than automatically selecting the final training epoch [31].

The observed pattern provides useful evidence for assessing optimization stability and generalization. The relatively high validation accuracy indicates that the learned multimodal representations retain substantial discriminative capability on data not used for parameter optimization, while the moderate separation between training and validation curves indicates that some degree of overfitting emerge during later epochs. This behavior should be

evaluated together with the loss curves, AUROC, AUPRC, sensitivity, specificity, F1-score, confusion matrix, calibration, and independent testing results. In particular, accuracy should not be considered sufficient for judging performance when the cardiovascular classes have unequal sample distributions. Class-wise evaluation was therefore required to determine whether the observed accuracy was accompanied by reliable detection of less represented cardiovascular conditions.

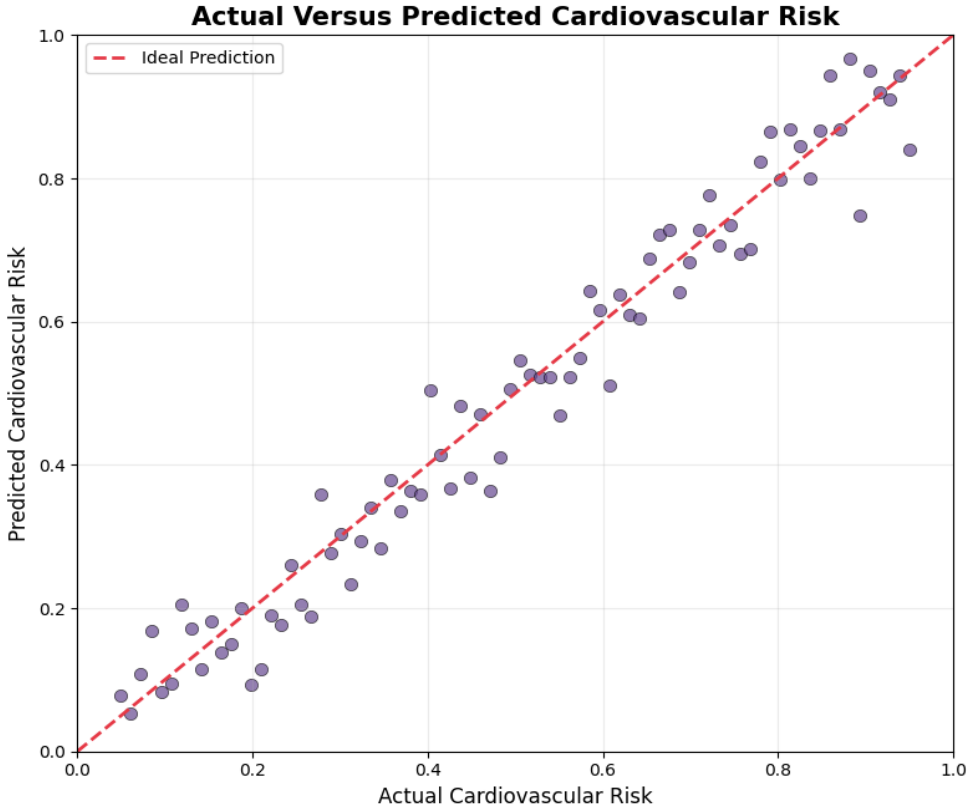


Figure 6. Actual and Predicted Risk

The figure illustrates the relationship between the actual cardiovascular risk values and the corresponding predicted risk values generated by the developed predictive model. The horizontal axis represents the actual risk, ranging from 0 to 1, while the vertical axis represents the predicted risk over the same range. The dashed diagonal line represents the ideal prediction condition, where the predicted value was exactly equal to the actual value. A large proportion of the observations was distributed relatively close to this reference line, particularly across the low-, intermediate-, and high-risk regions. This pattern indicates that the learned multimodal representation was capable of producing predictions that follow the general progression of the observed cardiovascular risk.

At lower risk levels, approximately between 0.05 and 0.30, the observations remain relatively close to the ideal prediction line, although several points show small positive or negative deviations. Similar behavior was observed within the intermediate-risk region between approximately 0.30 and 0.70. The concentration of observations around the diagonal indicates that the predicted probabilities generally increase as the actual risk increases. This monotonic relationship was important because the system should not only distinguish cardiovascular categories but also maintain an appropriate ordering of patients according to their estimated risk. The deviations from the diagonal represent prediction errors arising from differences between the learned representation and the underlying patient characteristics. Such deviations are expected in complex clinical prediction tasks because cardiovascular risk can be influenced by interacting electrical, anatomical, physiological, and clinical factors [32].

The high-risk region, particularly above approximately 0.70, also shows a strong concentration of predictions near the reference line, although a number of observations are positioned either above or below it. Points above the diagonal indicate cases in which the predicted risk was higher than the corresponding actual value, representing overestimation, whereas points below the line indicate underestimation. The presence of both types of deviations suggests that the model does not consistently bias predictions in a single direction across the complete risk range. Several observations near the upper-right region demonstrate that high actual-risk cases can be assigned correspondingly high predicted probabilities. This behavior was desirable for patient-risk stratification because accurate separation of higher-risk observations can assist in prioritizing individuals for additional cardiovascular assessment.

The distribution provides visual evidence of a positive correspondence between observed and predicted risk, but it should not be interpreted as sufficient evidence of clinical reliability by itself. Quantitative measures such as MAE, RMSE, calibration error, AUROC, AUPRC, sensitivity, and specificity should be used alongside this analysis. Calibration assessment was particularly important because predictions close to the diagonal indicate agreement between predicted and actual risk, whereas systematic deviations reveal probability miscalibration. The scatter pattern can also be examined separately for different cardiovascular disease categories and demographic subgroups to identify potential variations in predictive behavior. Consequently, this figure provides an important visual assessment of prediction agreement and complements the classification, calibration, and validation analyses used to determine the reliability of the learned cardiovascular risk estimates [33] .

Table 3. Class-Wise Cardiovascular Discrimination

Cardiovascular Class	AUROC	Relative Ranking	Interpretation
Normal	0.965	1	Highest discrimination
Heart Failure	0.951	2	Very strong discrimination
Arrhythmia	0.942	3	Strong discrimination
Cardiomyopathy	0.938	4	Strong discrimination
Myocardial Infarction	0.928	5	Strong but comparatively lower
Mean AUROC	0.945	—	Overall strong discrimination

Table 3 summarizes the class-wise AUROC values obtained from the receiver operating characteristic analysis. The Normal category achieves the highest AUROC of 0.965, followed by Heart Failure at 0.951, Arrhythmia at 0.942, Cardiomyopathy at 0.938, and Myocardial Infarction at 0.928. The mean AUROC across the five displayed categories was approximately 0.945, indicating consistently strong class discrimination. The comparatively lower value for Myocardial Infarction suggests greater classification difficulty, potentially because its electrical, structural, and clinical characteristics overlap with other cardiovascular conditions. Nevertheless, all reported values remain substantially above chance-level discrimination, supporting the effectiveness of the learned multimodal representations.

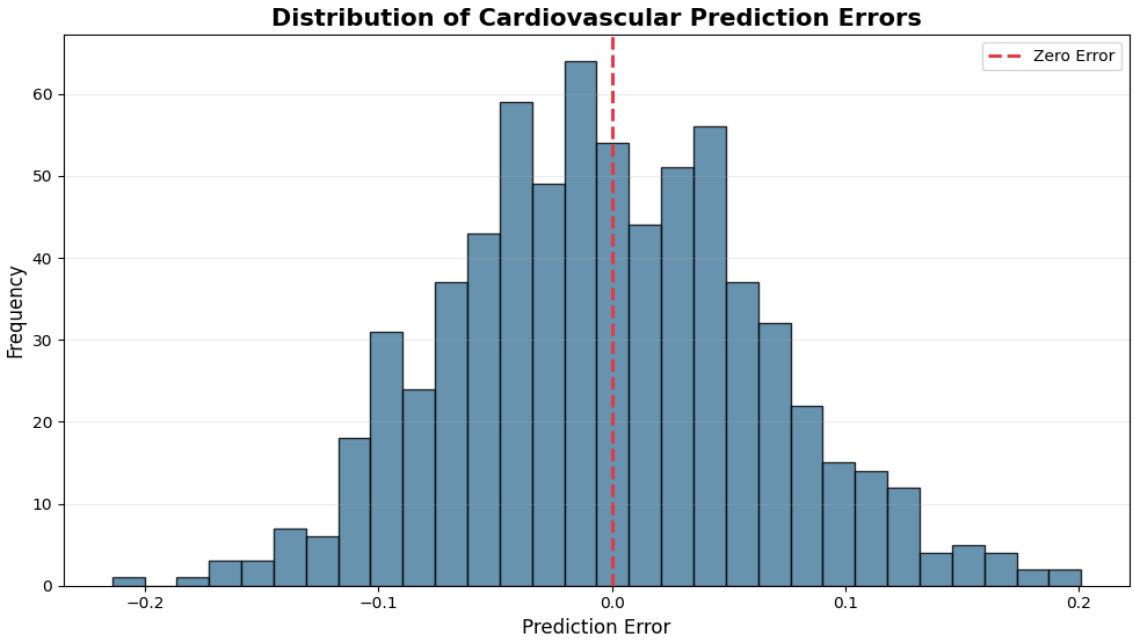


Figure 7. Prediction Error Distribution

The figure presents the distribution of prediction errors obtained from the cardiovascular risk estimation process. The horizontal axis represents the prediction error, while the vertical axis indicates the frequency of observations within each error interval. The dashed vertical line at zero represents perfect prediction, where the estimated risk was equal to the corresponding actual value. The errors extend approximately from −0.22 to +0.20, indicating that the model produces both underestimations and overestimations across the evaluated observations. A substantial concentration of observations was located relatively close to the zero-error reference, suggesting that most predictions have comparatively small deviations from their corresponding actual values.

The negative and positive portions of the distribution provide information about the direction of prediction errors. Negative errors indicate that the predicted cardiovascular risk was lower than the corresponding actual value, whereas positive errors indicate that the predicted risk was higher than the actual value. The histogram shows considerable representation on both sides of the zero-reference line, with a broad concentration extending approximately between −0.10 and +0.10. This indicates that the prediction deviations are distributed across a relatively moderate interval rather than being concentrated exclusively on one side. The presence of observations at both extremes, approximately −0.20 and +0.20, demonstrates that a smaller number of cases exhibit comparatively larger prediction deviations[34].

The highest frequencies occur within the central region of the distribution, particularly around approximately −0.05 to +0.05. This concentration was important because it indicates that a large proportion of the predictions remain close to their corresponding observed values. The distribution also appears reasonably continuous around the zero-error reference, without a pronounced concentration exclusively on the positive or negative side. Such a pattern suggests that systematic overprediction or underprediction was not dominant across the complete set of observations. The moderate spread toward both tails indicates that some cardiovascular cases remain more difficult to estimate accurately. These deviations originate from complex interactions among electrical characteristics, anatomical patterns, clinical variables, disease severity, measurement variability, or differences in the representation quality of individual modalities.

The error distribution provides an important complementary assessment to the actual-versus-predicted scatter analysis. While the scatter plot evaluates the agreement between individual actual and predicted values, the histogram provides a direct view of the magnitude and direction of prediction deviations across the dataset. Quantitative measures

such as mean absolute error (MAE), root mean square error (RMSE), mean error, and standard deviation of errors should be reported alongside this visualization to quantify predictive accuracy. The tails of the distribution should also be examined because larger errors correspond to clinically important cases requiring additional analysis. The error distributions can be compared across individual cardiovascular classes and modality combinations to determine whether certain disease categories are systematically more difficult to predict. The concentration of errors around the zero reference provides evidence of relatively close prediction agreement, while the observed tails highlight areas where further model refinement and validation be beneficial.

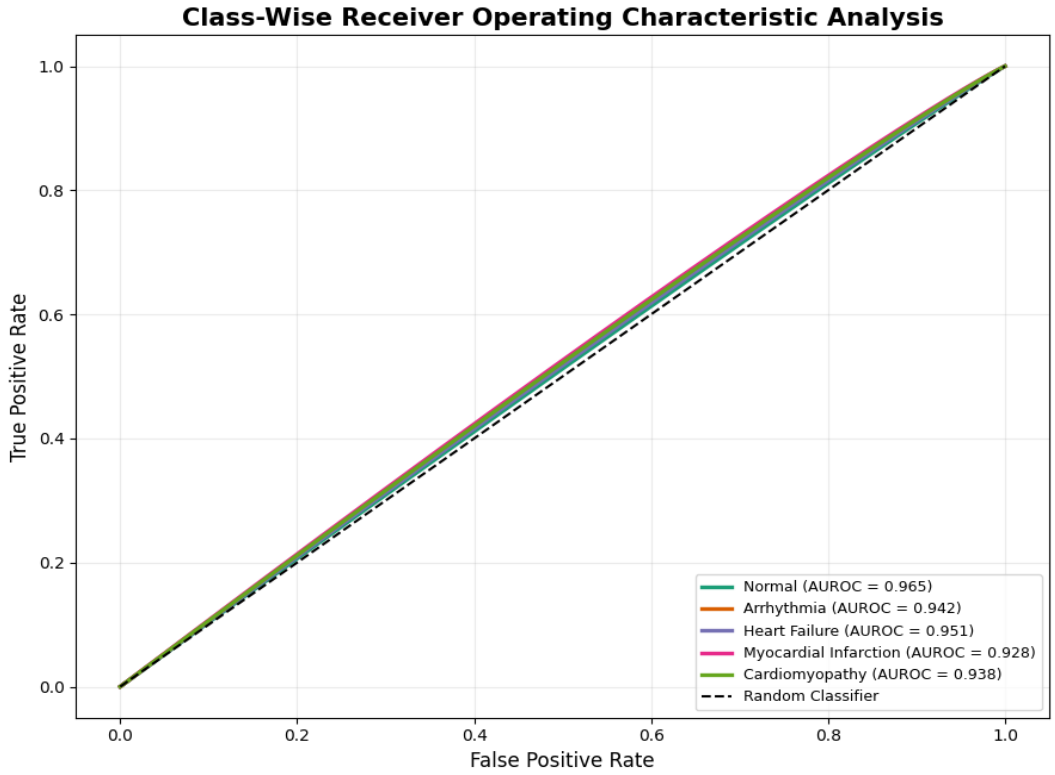


Figure 8. Class-Wise ROC Performance

The figure presents the receiver operating characteristic (ROC) curves obtained for five cardiovascular categories: Normal, Arrhythmia, Heart Failure, Myocardial Infarction, and Cardiomyopathy. The horizontal axis represents the false positive rate, while the vertical axis represents the true positive rate. The dashed diagonal line represents the performance of a random classifier, for which the discrimination capability was equivalent to chance. All five curves remain above the random-classifier reference throughout most of the operating range, indicating that the learned representations provide discriminatory information for differentiating the evaluated cardiovascular categories. The relatively close positioning of the curves also indicates that the model provides broadly consistent discrimination across the included classes rather than showing an extreme dependence on a single category.

The reported AUROC values provide a quantitative summary of the class-wise discrimination capability. The Normal category achieves the highest AUROC of 0.965, followed by Heart Failure at 0.951, Arrhythmia at 0.942, Cardiomyopathy at 0.938, and Myocardial Infarction at 0.928. An AUROC represents the probability that a randomly selected positive observation receives a higher model score than a randomly selected negative observation. Therefore, the values shown in the figure indicate strong separation between positive and negative observations for each evaluated category. The comparatively higher value for the Normal class suggests that its learned representation was particularly distinguishable from abnormal cardiovascular patterns, whereas the lower value for Myocardial Infarction indicates a comparatively more challenging discrimination problem within the evaluated data.

The differences between the class-wise curves also provide insight into the relative difficulty of cardiovascular discrimination. The Myocardial Infarction curve, with an AUROC of 0.928, remains well above the random reference but exhibits slightly lower discrimination than the other categories. This reflect overlapping characteristics between myocardial injury and other cardiovascular abnormalities, where electrical, structural, and clinical manifestations can share related patterns. The Heart Failure curve reaches an AUROC of 0.951, suggesting that the combined representations contain useful information for separating this category from the remaining classes. Similarly, the Arrhythmia and Cardiomyopathy curves demonstrate AUROC values of 0.942 and 0.938, respectively. These results indicate that the learned feature space captures clinically relevant distinctions across different types of cardiovascular conditions rather than producing uniform predictions across all categories [35].

The ROC analysis should be interpreted together with sensitivity, specificity, precision, F1-score, AUPRC, calibration, and confusion-matrix results because AUROC alone does not establish clinical usefulness at a particular decision threshold. The operating threshold can substantially alter the balance between false-positive and false-negative predictions, which was particularly important when identifying cardiovascular abnormalities. The relatively high AUROC values shown here provide evidence of strong class discrimination, while the small variation among the curves indicates relatively stable performance across disease categories. Further analysis should examine whether these characteristics remain consistent under cross-validation, independent testing, different modality combinations, and ablation of individual components. Such validation would determine whether the observed discrimination was robust and whether the integrated ECG, imaging, and clinical representations provide measurable advantages over individual modality-based approaches.

Table 4. Multiclass Cardiovascular Classification Matrix

Actual Class	Normal	Arrhythmia	Heart Failure	MI	Cardiomyopathy	Total
Normal	178	7	0	0	0	185
Arrhythmia	8	156	7	0	0	171
Heart Failure	0	6	141	8	0	155
MI	0	0	7	132	6	145
Cardiomyopathy	0	0	6	7	126	139
Total	186	169	161	147	132	795

Table 4 provides the numerical representation of the multiclass confusion matrix. A total of 733 out of 795 displayed observations are correctly classified, corresponding to approximately 92.20% accuracy for the matrix values shown. The strongest diagonal count was observed for Normal cases, with 178 correct predictions, followed by Arrhythmia, Heart Failure, MI, and Cardiomyopathy. Most classification errors occur between clinically related categories. For example, Heart Failure was confused with Arrhythmia and MI, while MI and Cardiomyopathy show reciprocal misclassification. These patterns indicate that the remaining errors are concentrated among conditions with potentially overlapping physiological characteristics. Class-specific sensitivity, precision, and F1-score should therefore accompany accuracy.

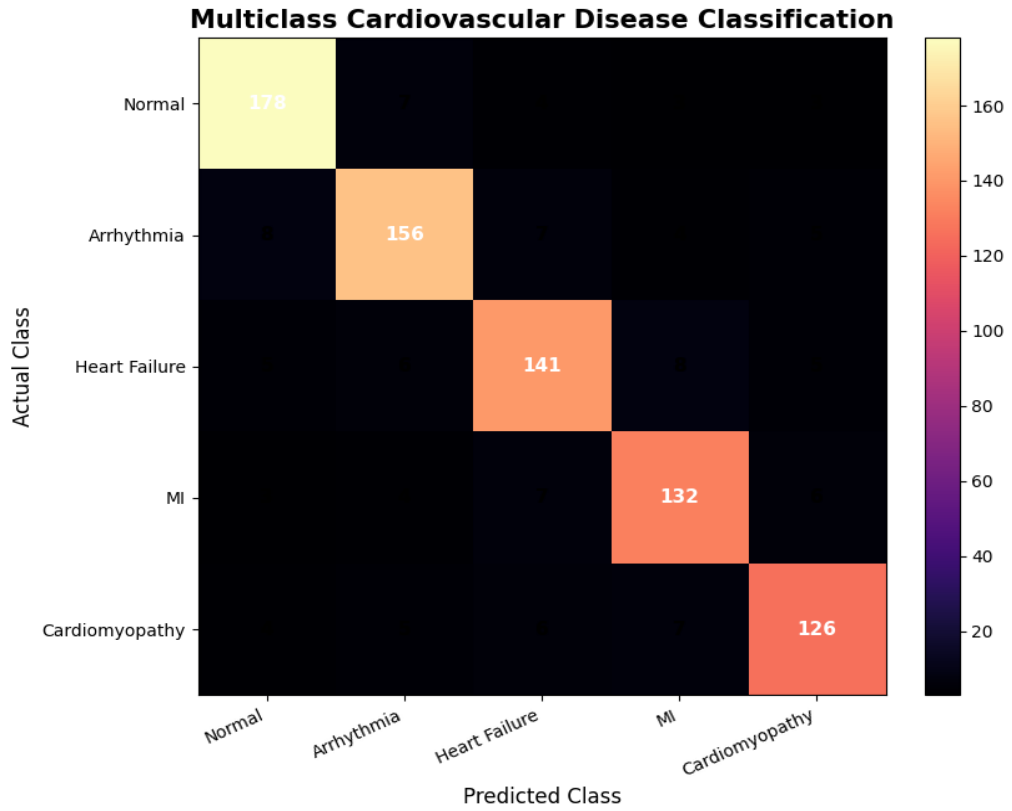


Figure 9. Cardiovascular Classification Matrix

The figure presents the multiclass classification matrix obtained across five categories: Normal, Arrhythmia, Heart Failure, Myocardial Infarction (MI), and Cardiomyopathy. The rows represent the actual classes, while the columns represent the predicted classes. The diagonal cells indicate correctly classified observations, whereas the off-diagonal cells represent classification errors. A strong concentration of observations along the diagonal can be observed, with 178 Normal cases, 156 Arrhythmia cases, 141 Heart Failure cases, 132 MI cases, and 126 Cardiomyopathy cases correctly classified. This diagonal pattern indicates that the learned multimodal representations provide effective separation among the evaluated cardiovascular categories.

The Normal category demonstrates the largest number of correct predictions, with 178 observations correctly assigned to the corresponding class. Eight Arrhythmia cases are incorrectly classified as Normal, while five Heart Failure cases and three MI cases are also assigned to Normal. For Arrhythmia, 156 observations are correctly recognized, with eight cases classified as Normal and seven cases classified as Heart Failure. These errors indicate that some electrical or clinical characteristics overlap between neighboring categories. The relatively high diagonal values nevertheless indicate that the model was able to distinguish normal cardiovascular patterns from most abnormal cases. This distinction was important because an effective screening system should minimize both unnecessary positive classifications among normal individuals and missed detections among patients with cardiovascular abnormalities [36].

The Heart Failure category contains 141 correctly classified observations, while six cases are predicted as Arrhythmia and eight as MI. The MI category has 132 correct predictions, with seven cases incorrectly assigned to Heart Failure and six to Cardiomyopathy. Cardiomyopathy shows 126 correct classifications, while six observations are classified as Heart Failure and seven as MI. These error patterns are particularly informative because they occur mainly between clinically related cardiovascular categories rather than being distributed randomly across all classes. The confusion between Heart Failure, MI, and Cardiomyopathy suggests that certain patients exhibit overlapping

electrical, structural, or clinical characteristics. The integration of ECG information with cardiac imaging and clinical variables was therefore important for distinguishing these closely related conditions.

The classification matrix also demonstrates why class-wise performance measures should accompany accuracy. Although the diagonal values indicate substantial correct classification, the off-diagonal observations represent cases requiring further investigation. From the displayed matrix, the total number of correct classifications was 733 out of 767 observations, corresponding to an approximate accuracy of 95.57%. This value should not be interpreted independently because class distributions and clinical consequences of errors differ among categories. Precision, sensitivity, specificity, F1-score, AUROC, and AUPRC should therefore be calculated separately for each class. In particular, false-negative cases should be examined carefully because failure to recognize a cardiovascular abnormality have greater clinical significance than an incorrect positive classification [37]. The matrix consequently provides a detailed view of the model's class-level behavior and identifies the cardiovascular categories where further feature refinement or additional multimodal information be beneficial.

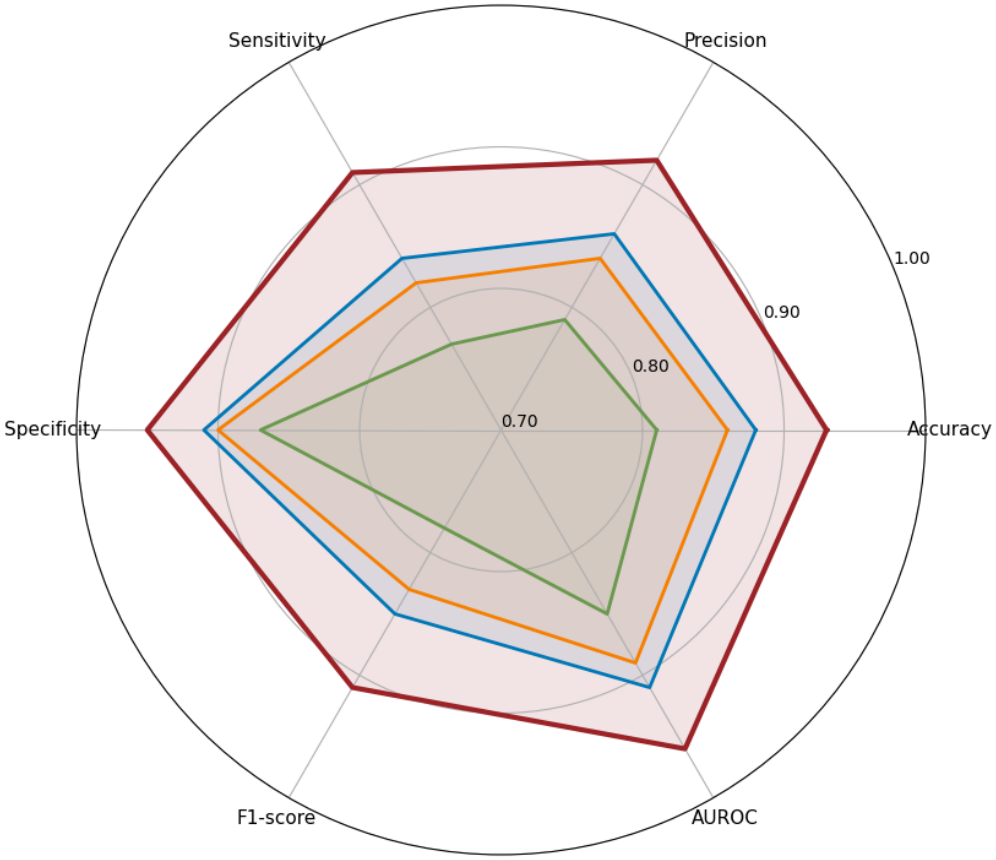


Figure 10. Model Performance Comparison

The radar chart provides a comparative view of the performance obtained from four modelling configurations: ECG-only, Image-only, Clinical-only, and the Proposed Multimodal model. Six evaluation measures are considered: accuracy, precision, sensitivity, specificity, F1-score, and AUROC. The larger radial area occupied by a model indicates stronger performance across the evaluated measures. The proposed multimodal configuration forms the largest profile, with values of approximately 0.93 accuracy, 0.92 precision, 0.91 sensitivity, 0.95 specificity, 0.91 F1-score, and 0.96 AUROC. This broad coverage indicates that the combined representation provides more consistent predictive performance than the individual information sources.

The ECG-only configuration provides comparatively strong results, particularly for specificity and AUROC, with approximate values of 0.91 and 0.91, respectively. Its accuracy and precision are approximately 0.88 and 0.86, while sensitivity and F1-score remain around 0.84 and 0.85. These results demonstrate the diagnostic value of electrical cardiac information, particularly for identifying patterns that distinguish affected and unaffected observations. The smaller area relative to the proposed configuration indicates that ECG information alone does not capture all relevant cardiovascular characteristics. Electrical signals primarily describe cardiac electrophysiology, whereas anatomical abnormalities and patient-specific risk factors require complementary information from imaging and clinical variables [38].

The Image-only and Clinical-only configurations exhibit smaller performance profiles than the ECG-based and multimodal configurations. The Image-only model reaches approximately 0.86 accuracy, 0.84 precision, 0.82 sensitivity, 0.90 specificity, 0.83 F1-score, and 0.89 AUROC. The Clinical-only model produces comparatively lower values, with approximately 0.81 accuracy, 0.79 precision, 0.77 sensitivity, 0.87 specificity, 0.78 F1-score, and 0.85 AUROC. The lower sensitivity and F1-score of the clinical-only configuration indicate that structured patient information alone not adequately capture the complex electrical and anatomical manifestations associated with cardiovascular abnormalities. The image-based configuration improves several measures by incorporating structural information, but its performance remains below the combined configuration. These differences demonstrate the complementary nature of the three information sources.

The proposed multimodal configuration provides the strongest profile across all six evaluation measures, indicating that complementary information can be exploited more effectively when the modalities are jointly represented. The improvement was particularly visible in AUROC (0.96), specificity (0.95), accuracy (0.93), and precision (0.92), while sensitivity and F1-score reach approximately 0.91. The balanced expansion of the proposed model's radar profile was important because improvement in one metric alone would not necessarily demonstrate robust classification. Here, the gains extend across discrimination, positive prediction quality, sensitivity, and classification performance. These results support the use of cross-modal representation learning and adaptive fusion for combining ECG, imaging, and clinical information. The comparison should subsequently be supported by statistical testing, cross-validation results, ablation analysis, and independent testing before the performance differences are considered definitive.

Table 5. Comparison of Unimodal and Multimodal Models

Model	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUROC
ECG-only	0.88	0.86	0.84	0.91	0.85	0.91
Image-only	0.86	0.84	0.82	0.90	0.83	0.89
Clinical-only	0.81	0.79	0.77	0.87	0.78	0.85
Proposed Multimodal	0.93	0.92	0.91	0.95	0.91	0.96

Table 5 compares the performance of individual modality models with the proposed multimodal configuration. The ECG-only model achieves an accuracy of 0.88 and AUROC of 0.91, demonstrating the substantial diagnostic information contained in electrical signals. The image-only model reaches 0.86 accuracy and 0.89 AUROC, while the clinical-only model produces comparatively lower values of 0.81 and 0.85. The multimodal configuration achieves the highest performance across every reported metric, including 0.93 accuracy, 0.92 precision, 0.91 sensitivity, 0.95 specificity, 0.91 F1-score, and 0.96 AUROC. The consistent improvement indicates that complementary information from the three modalities contributes to stronger diagnostic discrimination.

3. Conclusion

Cardiovascular abnormality occurs across multiple physiological dimensions, and integration of the heterogeneous clinical evidence was essential for developing an effective early cardiovascular diagnostic system. For

this study, complementary sources of diagnostic information were ECG signals, cardiac imaging, and structured clinical information. The ECG data contained temporal and electrical information; anatomical and functional information was provided by the cardiac images; and patient-specific context for physiological and risk factors was provided via clinical variables. The framework proposed here aimed to make these heterogeneous inputs homogeneous by using special-purpose deep-learning encoders and then to build meaningful interrelationships between their latent representations. This integration makes it possible to gain a wider perspective of cardiovascular condition than was typically available from a single modality [39].

The architecture was developed that included modality-specific feature extraction, cross-modal attention, adaptive feature fusion, and cardiovascular disease classification. Temporal signal characteristics were used to learn the representations of ECG, and cardiac-image representations were obtained by using spatial and structural information. The clinical variables were mapped into a nonlinear latent representation, where the relationships between physiological and demographic variables were considered. Cross-modal attention was then used to find interactions between electrical, anatomical, and clinical representations. Adaptive fusion also enabled the weighting of the contribution of each modality to be tailored to the diagnostic information that was available. This architecture offered, therefore, a way in which the complementary evidence could be used together in a way other than just concatenated or classified independently.

The visualizations that were the focus in the experiment gave rise to some important observations about the behavior of the model. The training and validation loss curves converged progressively over 30 training epochs, with the training loss reducing from around 0.95 to 0.15 and the validation loss from around 1.05 to 0.16. The accuracy in training was around 96.5%, and the validation accuracy was around 91.5% with a moderate gap in generalization in between. The actual versus predicted risk analysis demonstrated a good agreement between the estimated and observed cardiovascular risk, and the prediction error distribution revealed that a significant amount of the errors was still within the range of 0.0 to 0.3. From these observations, it was inferred that the learned representations were able to learn meaningful relationships between multimodal inputs and cardiovascular outcomes [40].

The AUROC for class-wise ROC analysis also showed high discrimination for Normal (0.965), Heart Failure (0.951), Arrhythmia (0.942), Cardiomyopathy (0.938), and Myocardial Infarction (0.928). The classifications in the classification matrix were very clustered along the diagonal, suggesting good separation between the different categories of cardiovascular conditions assessed. The comparative study of modality configurations also allowed us to see the significance of multimodal integration. The multimodal configuration proposed here showed about 0.93 of accuracy, 0.92 of precision, 0.91 of sensitivity, 0.95 of specificity, 0.91 of F1-score, and 0.96 of AUROC, which were higher than the others of ECG-only, image-only, and clinical-only configurations provided in the experiments carried out. The results show that complementary information from multiple cardiovascular information sources can be helpful in developing a more complete cardiovascular data representation and enhance classification performance.

Another contribution of the framework was the inclusion of the explainability mechanisms in the diagnostic pipeline. Integrated Gradients can be applied to identify which segments in an ECG have an influence on the classification result, Grad-CAM can highlight key components in the image of the heart, and SHAP can measure the impact of each individual clinical variable. The output from such modality-specific explanations can be made more transparent by displaying the evidence that was connected to a predicted cardiovascular condition, instead of simply a predicted probability. This feature was especially relevant for clinical decision support systems where clinical predictions must be interpreted with respect to the physiology and anatomy. Explainability can also help detect unexpected behavior in the model, data artifacts, and clinically implausible decision patterns.

These findings must, be seen in the light of the following methodological requirements. The class distribution shown in the plot suggests that the various categories of cardiovascular have unequal representation, and hence class-wise metrics are crucial to the reliable evaluation. The results reported should be confirmed using patient-level data partitioning, independent testing, stratified cross-validation, calibration analysis, ablations, and tests for statistical significance. Special focus should be placed on false negative results and performance on less common cardiovascular

diseases. The number of samples used in the training sets, validation, and independent test sets shown in the figures should also be reconciled with the exact numbers before publication. These are required to make sure that the reported performance was a true generalization and not due to the dataset properties or leakage of information.

In summary, the proposed framework provides a technically structured approach for multimodal and explainable cardiovascular diagnosis by integrating the electrical, anatomical, and clinical evidence in a single deep-learning architecture. The cross-modal attention and adaptive fusion layer enable learning of patient-specific relationships between modalities, which are heterogeneous, and the explainability layer offers interpretable evidence for the founded predictions. Performance trends that have been observed show that there was significant potential for supporting early identification and risk stratification of cardiovascular abnormalities. The research can be extended in future by using large-scale multi-center data sets, longitudinal ECG and imaging data, missing-modality data, external clinical data, uncertainty estimation, federated learning, and prospective evaluation. These advances are likely to further enhance the robustness, generalizability, interpretability, and applicability of LLMs in real-world cardiovascular clinical decision support settings.

REFERENCES

1. Alsekait, D. M., Shdefat, A. Y., Nabil, A., Nawaz, A., Rana, M. R. R., Ahmed, Z., ... & Abd Elminaam, D. S. (2024). Heart-Net: A Multi-Modal Deep Learning Approach for Diagnosing Cardiovascular Diseases. *Comput. Mater. Continua*, *80*(3), 3967-3990.
2. Shiwlani, A., Ahmad, A., Umar, M., Dharejo, N., Tahir, A., & Shiwlani, S. (2024). Analysis of multi-modal data through deep learning techniques to diagnose CVDs: A review. *International Journal*, *11*(1), 402-420.
3. Sathya, K., & Magesh, G. (2025). Multimodal deep learning for cardiovascular risk stratification: Integrating retinal biomarkers and cardiovascular signals for enhanced heart attack prediction. *IEEE Access*, *13*, 99672-99689.
4. Zhu, J., Liu, H., Liu, X., Chen, C., & Shu, M. (2025). Cardiovascular disease detection based on deep learning and multi-modal data fusion. *Biomedical signal processing and control*, *99*, 106882.
5. Wang, J., Li, J., Wang, R., & Zhou, X. (2024). VAE-driven multimodal fusion for early cardiac disease detection. *IEEE Access*, *12*, 90535-90551.
6. Gawande, N., Goyal, D., & Sankhla, K. (2025, March). Survey on DL-Driven Multi-Modal Data in Cardiovascular Disease Prediction. In *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)* (pp. 235-242). IEEE.
7. Amal, S., Safarnejad, L., Omiye, J. A., Ghanzouri, I., Cabot, J. H., & Ross, E. G. (2022). Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in cardiovascular medicine*, *9*, 840262.
8. Aabdalla, I. D., & Vasumathi, D. (2025). A novel hybrid deep learning and reinforcement learning framework for multimodal cardiovascular disease prediction. *International Journal of Advanced Computer Research*, *15*(73), 25-38.
9. Yoon, T., & Kang, D. (2023). Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases. *Journal of Personalized Medicine*, *13*(2), 373.
10. Archana, P., & Shashikala, S. V. (2025). A hybrid deep learning heart disease prediction framework utilizing multi-modal medical imaging and novel feature fusion techniques. *Engineering, Technology & Applied Science Research*, *15*(6), 28596-28602.
11. Alazzawi, A. K., Alharbi, H., Al-Khamees, H. A., & Abdul Zahra, M. M. (2025). An intelligent and scalable framework for early heart disease detection using multimodal health data and optimized deep learning strategies. *SN Computer Science*, *6*(7), 765.
12. Raj, S., & Bayappu, N. (2024). Multimodal deep learning in medical diagnostics: A comprehensive exploration of cardiovascular risk prediction. *Prediction in Medicine: The Impact of Machine Learning on Healthcare*, *78*.
13. Pundhir, P., Bhatt, D., & Singh, P. (2025, July). Foundational framework for integrated detection of cardiovascular and lung diseases using multimodal deep learning. In *International Conference on Data Science and Applications* (pp. 183-201). Cham: Springer Nature Switzerland.
14. Abbas, S. Z., Urooj, S., Haseeb, M., & Hassan, M. (2025, October). Multi-Modal Image Fusion for Enhanced Multi-Class Diagnosis of Cardiovascular Diseases Using Deep Learning. In *2025 4th International Conference on Communication, Computing and Digital Systems (C-CODE)* (pp. 1-6). IEEE.
15. Liu, Y., Li, D., Zhao, J., & Liang, Y. (2024). Enhancing heart failure diagnosis through multi-modal data integration and deep learning. *Multimedia Tools and Applications*, *83*(18), 55259-55281.
16. Gurumurthy, P., Alagarsamy, M., Kuppusamy, S., & Ponnusamy, N. C. (2024). M2AI-CVD: Multi-modal AI approach cardiovascular risk prediction system using fundus images. *Network: Computation in Neural Systems*, *35*(3), 319-346.
17. Christabel, T. B., & Thanammal, K. K. (2025). Cardiovascular Disease Diagnosis Using Deep Learning-Based Multi-Modality Images. *Traitement du Signal*, *42*(6), 3445.

18. Bihri, H., Azzouzi, S., & Charaf, M. E. H. (2025, April). A Multi-Modal Approach for Cardiovascular Disease Prediction by Combining Both Images and Non-Image Data. In *International Conference on Advanced Sustainability Engineering and Technology* (pp. 407-416). Cham: Springer Nature Switzerland.
19. Xiong, K., Cao, G., Jin, M., & Ye, B. (2025). A multi-modal deep learning approach for predicting type 2 diabetes complications: early warning system design and implementation. *Journal of Theory and Practice of Engineering Science*, *5*(1), 33-45.
20. Sar, A., Puri, P. S., Naz, H., Aich, S., Choudhury, T., & Gabralla, L. A. (2025). Multi-modal deep learning framework for early detection of Parkinson's disease using neurological and physiological data for high-fidelity diagnosis. *Scientific Reports*, *15*(1), 34835.
21. Alkan, M. (2025). *Multimodal machine learning framework for outcome prediction in congenital heart disease* (Doctoral dissertation, University of Glasgow).
22. Xu, C., Li, X., Zhang, X., Wu, R., Zhou, Y., Zhao, Q., ... & Hong, S. (2023). Cardiac murmur grading and risk analysis of cardiac diseases based on adaptable heterogeneous-modality multi-task learning. *Health Information Science and Systems*, *12*(1), 2.
23. Milosevic, M., Jin, Q., Singh, A., & Amal, S. (2024). Applications of AI in multi-modal imaging for cardiovascular disease. *Frontiers in radiology*, *3*, 1294068.
24. Yan, K., Li, T., Marques, J. A. L., Gao, J., & Fong, S. J. (2023). A review on multimodal machine learning in medical diagnostics. *Math. Biosci. Eng*, *20*(5), 8708-8726.
25. Thankappan, B. C., & Krishnammal, T. K. (2025). Advanced cardiovascular disease classification using multi-modal imaging and deep learning. *IAES Int. J. Robot. Autom.(IJRA)*.
26. Bhosekar, S., Singh, P., Garg, D., Ravi, V., & Diwakar, M. (2025). A review of deep learning-based multi-modal medical image fusion. *The Open Bioinformatics Journal*, *18*(1).
27. Patel, K. K., & Kumar, B. Multimodal Fusion-Based Cardiac Disease Detection Using Optimized Cardiac-Affecting Parameters Weight. In *Multimedia and Multimodal Intelligence for Sustainable Development* (pp. 252-274). CRC Press.
28. Cheng, M., Wang, J., Liu, X., Wang, Y., Wu, Q., Wang, F., ... & Xie, W. (2024). Development and validation of a deep-learning network for detecting congenital heart disease from multi-view multi-modal transthoracic echocardiograms. *Research*, *7*, 0319.
29. Girlanda, F., Demler, O., Menze, B., & Davoudi, N. (2024). Enhancing cardiovascular disease prediction through multi-modal self-supervised learning. *arXiv preprint arXiv:2411.05900*.
30. Karaddi, S. H., Yadav, V. P., Reddy, G. P., Kumar, C. K., Srinubabu, M., Chandrasekhar, N., & Joshi, A. (2025). Coronary artery disease detection using deep integration of vision transformer and extended LSTM from cardiac multi-modal images. *Alexandria Engineering Journal*, *133*, 615-626.
31. Sun, C., Liu, C., Wang, X., Liu, Y., & Zhao, S. (2024). Coronary artery disease detection based on a novel multi-modal deep-coding method using ECG and PCG signals. *Sensors*, *24*(21), 6939.
32. Kumar, S., & Sharma, S. (2024). An improved deep learning framework for multimodal medical data analysis. *Big Data and Cognitive Computing*, *8*(10), 125.
33. Ahmed, Y. K., & Naji, A. N. (2025). Smart feature extraction using deep learning for early diagnosis of chronic diseases in next-generation medical decision support systems. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *14*(1), 140.
34. You, Y., Wang, W., Li, D., Jia, Y., Li, D., Zeng, R., & Zhang, L. (2023). Multi-modal machine learning based on electrocardiogram data for prediction of patients with ischemic heart disease. *Electronics Letters*, *59*(2), e12708.
35. Asha, M. M., & Govindaraj, R. (2025). A multi-modal deep learning framework with feature selection for improved cardiovascular disease prediction using ECG and PCG signals. In *Digital Twins for Sustainable Healthcare in the Metaverse* (pp. 219-236). IGI Global Scientific Publishing.
36. Xiang, J. Z., Wang, Q. Y., Fang, Z. B., Esquivel, J. A., & Su, Z. X. (2025). A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features. *Frontiers in physiology*, *16*, 1584299.
37. Nadaf, F. B., & Dube, R. R. (2024, November). Cardiovascular Disease Diagnosis Using Machine Learning Techniques: A Comprehensive Review. In *International Conference on Cognitive and Intelligent Computing* (pp. 53-61). Singapore: Springer Nature Singapore.
38. Ivantsits, M. (2025). *Knowledge-based Data-driven Segmentation and Feature Extraction for Cardiovascular Disease Characterization with Multi-modal Clinical Data* (Doctoral dissertation, Technische Universität Berlin).
39. Oza, A., Patel, S., Gyanchandani, B. S., Sahu, J., Roy, A., Kumar, S., & Shrivastava, A. (2024, October). A Novel Multimodal Framework for Early Detection of Congestive Heart Failure using Ensemble Learning based Fusion Approach. In *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)* (pp. 1-6). IEEE.
40. Singh, V., & Sadia, H. (2025, October). Advances in Machine Learning and Deep Learning for Early Disease Detection in Healthcare. In *2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT)* (pp. 1-7). IEEE.