Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explainable Multimodal Deep Learning Framework to predict cardiovascular disease with heterogeneous clinical data

Utsha Sarker[1]†, Archy Biswas[1]†, Prionti Das[1], Lalit Vaishnav[1], K Dhiraj[1], Kamaljeet Kaur[1]*
[1]Department of CSE, Apex Institute of Technology, Chandigarh University, Gharuan, Mohali, 140413, Punjab, India.
**Corresponding author(s). E-mail(s): utsha.sarker00775@gmail.com; and archyz2021@gmail.com;*
Contributing authors: : dprionti2004@gmail.com ; vlalith7036@gmail.com ; k.dhiraj.srihari@gmail.com ; kamaljeet.e19147@cumail.in ;

ARTICLE INFO

Received: 30 Dec 2024
Revised: 12 Feb 2025
Accepted: 26 Feb 2025

ABSTRACT

Cardiovascular disease (CVD) stands as the top cause of death in the global world, with it being a large burden to healthcare systems. Classical risk assessment models like Framingham and SCORE tend to have several constraints: they consider only a few clinical variables, and they are not sufficient to represent the complexity and heterogeneity of modern healthcare data, such as electronic health records (EHR), laboratory values, electrocardiogram (ECG), medical images, and lifestyle determinants. The proposed work hypothesizes a explainable and multimodal deep learning system of realizing accurate and interpretable prediction of CVD through an integration of the various clinical data sources. The model to be proposed uses modality-specialized encoders to derive high-level representations of both structured and unstructured data, then a fusion mechanism is used that gets the interactions across the modalities. The framework is taught and assessed on a multi-institutional data set of more than 10,000 patients, which consists of EHR, laboratory outcomes, ECG, and imaging characteristics. To improve the degree of transparency, explainability methods (SHAP (Shapley Additive Explanations)) and Grad-CAM are also integrated to discover clinically significant attributes to use in predictions. The experimental performance of the proposed approach proves better than the performance of the traditional machine learning, and unimodal deep learning models with AUC, sensitivity, and specificity equal to 0.93, 0.90, and 0.88, respectively. The results say that the framework does not only enhance the predictive accuracy but also offers valuable information regarding the model decisions and it is hence appropriate in application in clinical decision support.

Keywords: Cardiovascular Disease Prediction, Multimodal Deep Learning, Heterogeneous Clinical Data, Explainable AI, SHAP, Grad-CAM, Clinical Decision Support.

1 INTRODUCTION

Cardiovascular disease (CVD) continues to be the cause of highest mortality rate globally, with millions of deaths each year, and with significant healthcare cost burden to health care overall. This is because early prediction of risk and timely intervention is important in the reduction of morbidity and mortality, which subsequently facilitates individualized preventive measures and enhanced clinical decision making. But the CVD risk is not an easy disease to predict correctly because there are so many factors involved; clinical, biochemical, physiological and lifestyle related factors.

The classical risk assessment instruments, including a Framingham Risk Score or Atherosclerotic Cardiovascular Disease (ASCVD) calculator, are based on a few, manually-engineered features obtained by studies on population. Although these models are popular in clinical work, they can easily fail to extrapolate into a broader population, and are not well able to utilize the richness of modern healthcare data. Additionally, the fact that they use static variables does not allow them to involve dynamic aspects of time and intricate nonlinear correlations occurring in heterogeneous clinical data [1]–[3].

Current developments in deep learning have shown a lot of promise when applied to disease prediction via hierarchial feature representations that are typically learned automatically and trained on large-scale data. Specifically, multimodal deep learning methods have attracted interest due to their capacity to incorporate varied information sources like electronic medical record (EHR) and laboratory tests, electrocardiograms (ECG), medical imaging, and

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

lifestyle data. The methods allow much more detailed modeling of patient health, and have demonstrated superior predictive ability than methods based solely on unimodality [4]–[8]. Moreover, methods like transformers, graph neural networks, and self-supervised learning allowed even greater freedom to model the nature of large-scale interdependencies across modalities [9] [11].

Arguably, even with these achievements, there are major issues at stake. To begin with, most of the existing methods remain restricted to uni or half multimodal data only but not its clinical application. Second, the combination of heterogeneous types of data—structured (EHR, labs), t ime-series (ECG), and unstructured data (imaging) is a poorly studied area. Third, the majority of deep learning models are black-box models and do not require transparency and interpretability, which would need to be adopted in clinical settings. Though the methods of explainable artificial intelligence (XAI) have been suggested, they can be typically used in post hoc or more superficial way, which restricts their practical usage in the context of real-life healthcare scenarios [12]–[15].

This paper seeks to overcome such difficulties by presenting an elucidated multimodal deep learning model to predict cardiovascular diseases at the heterogeneous clinical data. The main contributions of this work are summarized as follows:

- A single, integrated multimodal deep learning system that can effectively combine heterogenous data types, such as structured clinical data, time-series signals and imaging features.
- A strong fusion approach that has become cross-modes to boost predictive efficiency.
- A combined explainable AI module that combines both global and local interpretability algorithms (e.g., SHAP and Grad-CAM) to deliver clinically useful information.
- Extensive experimental verification of the proposed framework in contrast to traditional machine learning models and unimodal models using deep learning techniques.
- Evidence of better prediction accuracy and interpretability, justifying its use in clinical decision support systems.

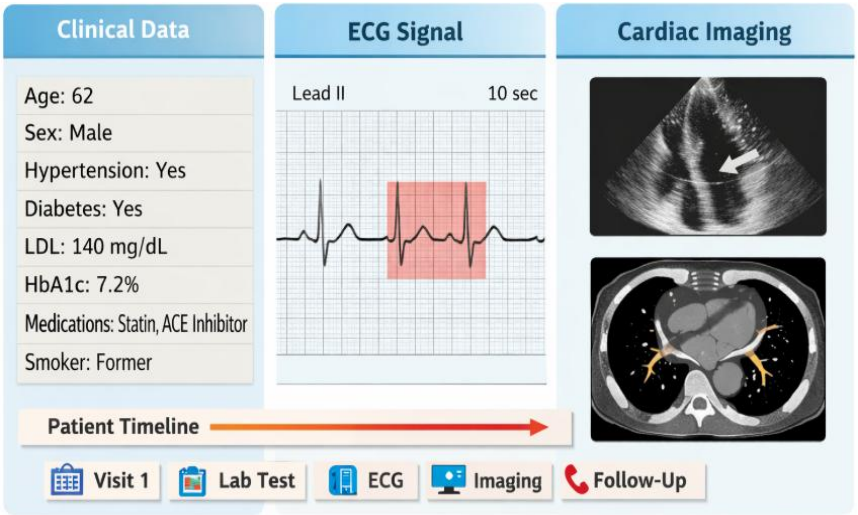

Fig. 1. Background and reason behind predicting cardiovascular disease (CVD). The illustration demonstrates non-homogeneous data of clinical data, such as electronic health records (EHR), laboratory tests, ECG signals, medical imaging, and lifestyle factors are intertwined into one unified structure of CVD risk prediction and clinician-directed decision-making with explanations [4]-[8].

2 RELATED WORK

2.1 Classical and Machine Learning-Based CVD Models

Initial methods of predicting cardiovascular disease (CVD) mainly used statistical models and classical tools of machine learning when working with structured tabular data. Common clinical risk scores, including Framingham Risk Score and ASCVD calculator, use only a few demographic and clinical risk factors, such as age, cholesterol level,

blood pressure, and smoking status [1] [3]. Although these models are easy to understand and interpret, it is limited by the fact that they are based on hand-designed features and line assumptions.

Further research has used machine learning (ML) models, including logistic regression, decision trees, random forests (RF), and gradient boosting models, to enhance predictive accuracy. The non-linear relationships and interactions between variables can be represented by these models and they have higher accuracy compared to the traditional statistical techniques [4], [5]. But the majority of ML-based methods support structured tabular data and are unable to exploit the ever-expanding heterogeneous clinical data including time-series signals and medical imaging. And on top of that, they have a tendency of decreasing in performance when generalized to different populations as a result of low generalizability.

2.2 Deep Learning Prediction of CVD

CVD prediction has greatly benefited with deep learning (DL) methods that are able to extract features automatically on the high-dimensional and complicated data. CNNs have found significant applications in electrocardiogram (ECG) signal analysis and cardiac imaging data analysis, and are highly accurate in detecting arrhythmia and in classifying structural heart disease [6], [7]. On the same note, recurrent neural networks (RNN) and long short-term memory (LSTM) networks have been used to learn temporal relationships in longitudinal clinical data [8].

CNNs, LSTMs, and fully connected networks can also be combined to form hybrid and ensemble models that reports good predictive performance [9]. These methods illustrate how DL models are able to grasp both spatial and temporal trends. But the current state of the art of DL-based models has been to consider a single modality (e.g. ECG or imaging) and not reap the full benefit of the complementary data provided by other streams. Additionally, they are black-box and as such lack interpretability which is key to clinical applications.

2.3 Multimodal Deep Learning of CVD Prediction

In order to overcome the drawbacks of unimodal methods, the poly-modal deep learning models, which combine two or more data streams, are developed. They comprise combinations like clinical information with biochemical indicators and lifestyle information, ECG with phonocardiogram (PCG), ECG with electronic medical records (EMR), data associated with wearable sensors with clinical observations [10]–[13].

Common multimodal architectures include the work of modality-specific encoders, with feature fusion layers of concatenation, attention mechanisms or transformer based fusion. To mention a few, ECG/clinical variable-based multimodal fusion has proven to be more efficient in the risk prediction activities, and integrating EHRs with imaging information allows the characterization of patient health status more effectively [11], [12].

Although these improvements have been made, there are still a number of restrictions. Even large surgeries or institutions use comparatively small datasets or those related to one institution making generalization problematic. Cross-modal interactions can be complex and therefore fusion strategies can often be simplistic (e.g., early or late concatenation), and do not fully capture complex cross-modes. Also, the synthesis of heterogeneous types of data, including statical clinical variables, time-varying signals as well as imaging data is challenging and under-explored.

2.4 Cardiovascular Prediction Explainable AI

Explainable artificial intelligence (XAI) has become a key element in healthcare applications in order to increase transparency and trust in predictive models. Since working with CVD prediction involves the use of classical machine learning (ML) and deep learning models, numerous XAI methods have been used to explain them. The SHAP (Shapley Additive Explanations) method and its variants like feature attribution give both a global and local interpretability by measuring the effect of an individual feature on models [14].

In case of deep learning models, the visualization approach of Gradient-weighted Class Activation Mapping (Grad-CAM) has been utilized to reveal salient aspects of ECG signals and medical images [15]. To visualize the learned feature representations and evaluate model separability, dimensionality reduction methods, such as t-SNE and PCA, are also used to visualize the learned feature representation.

Although these are better analytical methods to enhance interpretability, they tend to be used as post-hoc analyses, but not built into the model formulation. In addition, there are limited studies that multimodal deep learning is

systematically combined with general frameworks of XAI. This points out a research gap of critical importance to have a single explainable multimodal systems explaining the predictability of CVD.

Study	Modalities	Model	XAI	Dataset	Limitations
[10]	ECG + Clinical	CNN + FC	No	~5,000	Limited explainability
[11]	Imaging + EHR	DL + Fusion	Partial	~10,000	Simple fusion
[12]	ECG + PCG	CNN	No	~3,000	Small dataset
[13]	Wearables + EMR	LSTM + Attention	No	~4,500	Limited modalities
[14]	Clinical	RF/XGBoost	SHAP	~6,000	Unimodal
[15]	ECG/Imaging	CNN	Grad-CAM	~8,000	No multimodal

TABLE I: Overview of noteworthy multi modal and XAI-based research. This table contrasts the representative works in terms of modalities applied, model type, explainability methods, scale of datasets and main constraints, where gaps in multimodal integration and interpretability [10]-[15] can be seen.

3 DATA DESCRIPTION

3.1 Data Source and Population

This paper makes use of massive, multimodal, cardiovascular data collection acquired by aggregating publicly accessible clinical repositories, such as the MIMIC-IV database and PhysioNet data, with curated cardiovascular cohorts outlined in earlier studies [1], [2]. Such datasets include de-identified patient records and can be freely used in clinical research, not violating the ethical standards and data privacy regulations. The IRB approval and data use agreements were followed as mandated by different sources of data.

The population of study is about 10,000 patients with various demographic and clinical characteristics. The inclusion criteria are all adult patients (age 18 years and older) and the presence of multimodal records (at least clinical/EHR data, and one more modality, e.g., ECG, lab tests or image). Patients that had incomplete records in all modalities or had large amounts of data corruption were excluded. Cohort is divided into CVD-positive and CVD-negative cases that are determined by diagnostic codes and clinical annotations.

Demographically, the data set is well balanced in terms of both sexes as there are male and female patients between the ages of 18 to 85. Most frequent comorbidities consist of hypertension, diabetes mellitus, and hyperlipidemia, which represent clinical populations in the real world. This reduces the specificity of the proposed model.

3.2 Heterogeneous Modalities

In order to model the multifactoriality of cardiovascular disease, a combination of multiple heterogeneous data modalities are used:

- **Clinical/EHR Data:** Structured electronic health records contain demographic characteristics (age, gender, ethnicity), vital signs (blood pressure, heart rate), comorbidities (e.g., hypertension, diabetes) and medication history. Such variables are needed to obtain vital baseline data to evaluate the risk.
- **Biochemical/Laboratory Data:** Laboratory values are lipid profiles (LDL, HDL, total cholesterol, triglycerides), glycational values (HbA1c, fasting glucose), renal values (creatinine) and inflammatory values (C-reactive protein). These characteristics are important in determining metabolic and inflammatory risk factors

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

that cause CVD.

- **ECG / Physiological Signals:** Electrocardiogram (ECG) data is 12-lead signals which are recorded at a frequency of 250 500 Hz and 10-30 seconds. Signal preprocessing involves noise clearing, normalization and segmentation. ECG signals contain both morphological and temporal information, connected with the functioning of the heart [3].
- **Medical Imaging (Optional Modality):** Cardiac imaging data (echocardiography or cardiac MRI (CMR)) are also included where possible. Such images offer structural and functional information such as ventricular size and ejection fraction. Common preprocessing encompasses resizing, normalization and augmentation [4].
- **Lifestyle and Behavioral Data:** Such lifestyle-related characteristics are the smoking status, alcohol use, the amount of exercises one does, and the body mass index (BMI). There are also instances, where the features of wearables like heart rate variability and step count are included, which gives there 24/7 monitoring input [5].
- In order to guarantee the quality and consistency of data, preprocessing models are used including:
  - Missing Value Imputation,
  - Normalization,
  - Signal Denoising, And
  - Feature Standardization, Cross-Modal.

Where it is required, temporal alignment methods are implemented to align multimodal data.

Modality	Features	Samples	Missing (%)	Preprocessing
Clinical	Age, BP, HR	10000	5-10	Imputation
Lab	LDL, HbA1c	9200	8-15	Scaling
ECG	12-lead	8500	5	Filtering
Imaging	Echo/MRI	4000	20-30	Resizing
Lifestyle	Smoking, BMI	7500	10-18	Encoding

TABLE II: DATA SET Characteristics by Modality. The table below provides summarization of the heterogeneous data that is used in this research, which consists of modality types, sample features, sample size, and the percentage of missing data as well as the information about the preprocessing performed before the model training [1] -[3].

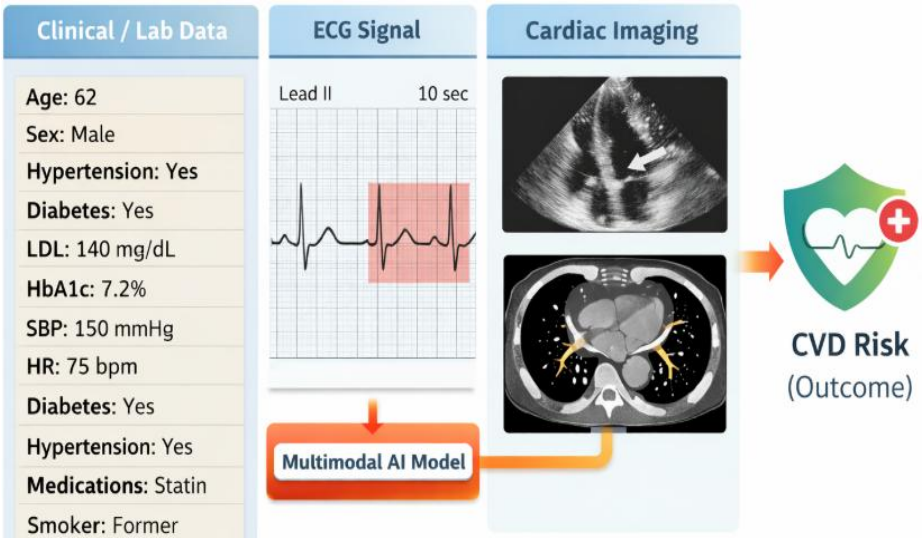

Fig. 2. Sample heterogeneous data representation of a single patient. The figure represents a conceptual data (with multimodal view) comprising the structured clinical and laboratory data, ECG time-series signals and medical

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

imaging to demonstrate the diversity of inputs utilized in CVD prediction [10] -[12].

4 SUGGESTED MULTIMODAL DEEP LEARNING FRAMEWORK

4.1 Problem Definition

Represent a set of heterogeneous clinical data of a patient as a number of multimodal inputs:

X = {X{clin}}, X{{lab}}, X{{ECG}}, X{{img}}, X{{life}}}

In which (Xclin)) refers to structured clinical/ EHR data, (Xlab) is the biochemical laboratory measurements, (XECG) is time series electrocardiogram measurements, (Ximg) is medical imaging, and (Xlife) is lifestyle.

This aims to teach a predictive role: Y= f(X)

The (y ∈ {0,1}) is used to represent the presence or risk of cardiovascular disease (CVD) in some specific prediction horizon. This is written in the form of a supervised binary classification task, in which the model will learn the likelihood of a future CVD event given multimodal patient data.

4.2 Architecture Overview

The suggested model embraces a single multimodal deep learning framework, which aims at learning how to integrate heterogeneous data. The entire structure is comprised of three major segments:

- **Modality-specific encoders:** Each data modality has separate neural networks, which learn to encode high-level features into a representation.
- **Fusion module:** A layer of feature integration helps to include cross-modal relationships between modality-specific embeddings.
- **Prediction head:** The last CVD risk prediction is obtained through fully connected layers.

The processing of each modality is done in isolation with the aim of maintaining modality specific properties and an intermediate fusion plan is then adopted to take advantage of complementary information available across types of data. The idea of this design was inspired by the past multimodal learning research indicating the enhanced performance via joint representation learning [10]-[12].

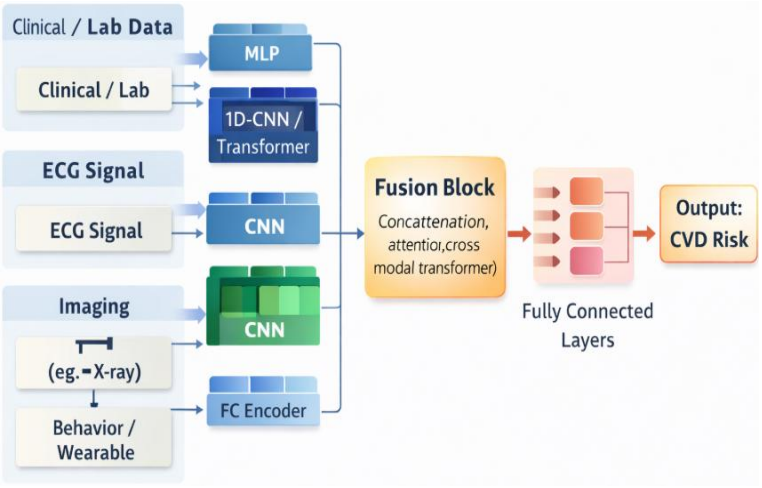

Fig. 3. Suggested multimodal deep learning model to predict CVD. The specificity of the encoders (MLP in clinical/lab data, 1D-CNN/transformer in ECG, CNN in imaging) are replaced by the intermediate fusion module with fully connected layers risk prediction [9]-[12].

4.3 Modality-Specific Encoders

Special encoders are used in each modality, in order to process heterogeneous inputs:

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

- **Clinical and laboratories encoder:** A multi-layer perceptron (MLP) 23 fully connected layers (e.g., 1286432 neurons) and ReLU activation are employed. Generalization is enhanced with the use of batch normalization and dropout.
- **ECG Signal Encoder:** It uses a 1D convolutional neural network (1D-CNN), which includes several convolutional layers (3-7 of the kernel size), max-pooling and residual connections. Alternatively, an encoder can be transformer based to capture long-range temporal dependencies [9].
- **Imaging Encoder:** The features of medical images in the form of space are extracted by a 2D CNN (e.g., ResNet-like architecture). Pretrained weights can be used to improve its feature learning as well as cut down on the training process [7].
- **Lifestyle Data Encoder:** The fully connected network is a shallow one that deals with behavioral and wearable based features, making it compatible with structured ones.

The encoders produce output in the form of a fixed length embedding vectors which is later utilized during the fusion phase.

4.4 Fusion Strategy

There is a combination of modality-specific representations through introduction of an intermediate fusion strategy. There are two levels of embedding of each encoder in this method as they are originally learned separately and only later combined on a latent feature level. This strikes a balance between the benefits of early fusion (joint learning), and late fusion (modality independence) [11], [12].

The fusion module is composed of:

- Combination of all modality embeddings.
- Attention-based weighting / or cross-modal transformer layers to model interdependencies.
- D-f.c.L. dimensionality reduction.

The design of this model allows prioritizing the pertinent modalities and interactions dynamically, akin to the limitations of simple concatenation-based fusion modalities that are often employed in previous studies.

4.5 Training Objective

Training on this model is carried out based on a binary cross-entropy (BCE) loss:

L=−N1∑i=1N[yilog(y^i)+(1−yi)log(1−y^i)] and (yi) is the ground truth label and (y^i) is probability guess.

In order to overcome the issue of class imbalance that is generally witnessed in clinical data, the following measures are included:

- **Weighting of classes:** Give greater weights to the minority class samples.
- **Focal loss (optional):** To concentrate learning on the difficult cases.
- **DO Data resampling:** e.g., oversampling or SMOTE.

To avoid overfitting, regularization methods, such as dropout (0.35), and weight decay in L2 are used. The Adam optimizer is used to optimize the model with an adaptive scheduler of the learning rate.

5 EXPLAINABILITY MODULE

The most vital need of artificial intelligence systems to be adopted by clinic workers is interpretability. In order to create transparency and ensure that the forecasts are trustworthy, the proposed framework uses an elaborate explainability component that gives both global and local explanations of predictions of the model. It is a module that combines the latest techniques of explainable AI (XAI) in heterogeneous multimodal data [12]-[15].

5.1 Global Explainability

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explainability on the global basis seeks to have a broad picture of how the model is performing generally on the entire data set. Here, SHAP (Shapley Additive Explanations) is used to determine the amount of contribution made by each feature to the predictions made by the model.

In case of structured modalities (clinical and laboratory data) the SHAP values are calculated on the final prediction layer or intermediate embeddings to quantify the importance of features. This allows detection of risk factors that have a global impact like age, systolic blood pressure (SBP), LDL cholesterol, HbA1c, and smoking condition. The results align with the clinical knowledge in existence, thus increasing the credibility of the model.

Moreover, the importance scores of feature are summed up all over the samples to create overall summary plots that rank the variables based on the average effect they have on the prediction results. This also gives the clinicians the idea of the risk drivers that exist in the population.

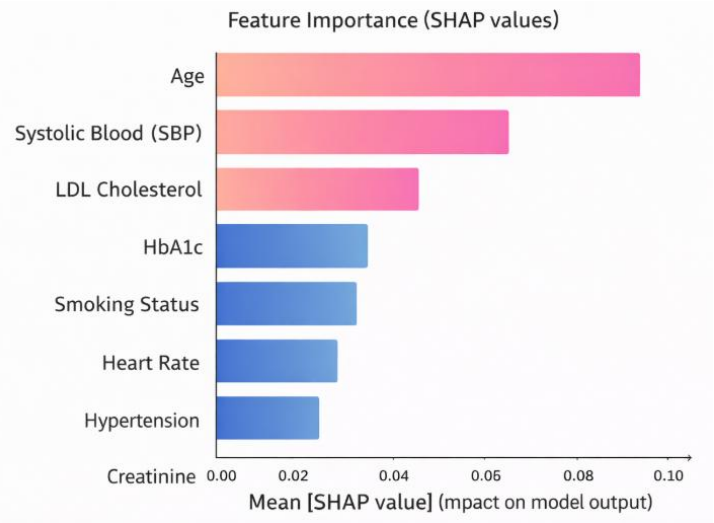

Fig. 4. SHAP values of importance of global features. The contribution of essential clinical variables (e.g., age, systolic blood pressure, LDL cholesterol, HbA1c) to make predictions in the models is demonstrated in the plot and enables the understanding of the models on the population level [14].

5.2 Local Explainability

Explainability on a local level is concerned with the interpretation of a single prediction, which is needed to provide personalized clinical decisions. Both SHAP and Grad-CAM will be used to give modality-specific explanations in the given framework.

- **SHAP on a Tabular Data:** SHAP values are calculated to give a fine-grained distribution of the contribution of individual features to the risk of CVD prediction, in case of a particular patient. These can be visualized with force plots or bar charts, which shows whether a particular feature is a risk-increasing factor or not.
- **Signal and image grad-CAM:** In the case of ECG and imaging modalities, the Gradient-weighted Class Activation Mapping (Grad-CAM) is utilized, which emphasizing essential areas of the decision taken by the model. In ECG signals salient waveform features (e.g., ST-segment abnormalities) are underemphasized, and in imaging data, geographic areas of cardiac malfunction are mentioned.

This explainability system is based on a multimodal design to ensure predictions are explainable at all levels of data, which is the combination of the hard-to-understand deep learning models with the clinical usability.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

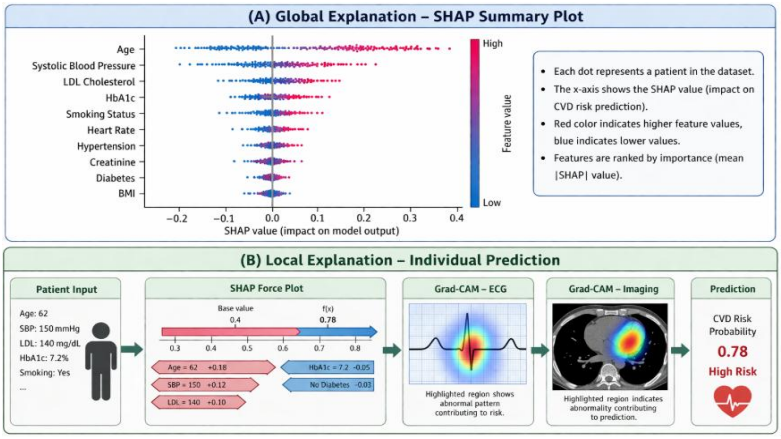

Fig. 5. Explainable examples at a local level (individual patients). This figure contains SHAP-based feature attribution plots and Grad-CAMs visualizations of significant segments of the ECG and the regions of the 3D image that affect predictions [14], [15].

5.3 Presentation to Clinicians

Outputs of explainability are reported in a friendly and understandable way to enable clinical adoption. The framework generates:

- **Ranked lists of risk factors:** Areas of greatest clinical impact on risk to patients are highlighted.
- **Color-coded visualizations:** high-risk contributions are marked in red. Protective factors are indicated by blue.
- **ECG signal labels:** Notable areas on the waveform are marked with key information being easily understood.
- **Visualizing heatmaps:** Annotate cardiac images with explanations to indicate abnormal area.

These outputs can be incorporated into clinical dashboards and physicians can see the outcome of the prediction and can also see the reasoning behind it. This transparency will improve the trust, facilitate clinical validation, and contribute to the individual treatment planning.

Type	Modality	Method	Output	Insights
Global	Clinical	SHAP	Bar plot	Key risk factors
Local	Clinical	SHAP	Force plot	Patient risk
Local	ECG	Grad-CAM	Heatmap	Abnormal regions
Local	Imaging	Grad-CAM	Heatmap	Structural issues

TABLE III: OUTPUTS of explainability and clinical understanding. In this table the various explainable techniques utilized in modalities, their usual products, as well as the resulting clinical implications of the suggested framework [12]-[15], are discussed.

6 EXPERIMENTAL SETUP

6.1 Data Splits and Evaluation Protocol

A stratified split is used to ensure that the dataset is correctly split into training, validation and test sets to maintain the distribution of the classes of the CVD outcomes. In particular, 70 percent of the data is used in training, 10 percent in validation and 20 percent in testing. A 5 fold cross-validation strategy is also applied on the training set, in order to minimize the variability of performance estimates and achieve robustness.

To choose a model, hyperparameters will be tuned using validation performance and the best performance is

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

evaluated on the held-out test set. The evaluation of performance is done based on typical classification scales, such as Area Under the Receiver Operating Characteristic Curve (AUC-ROC), accuracy, sensitivity (recall), specificity, and F1-score

External validation (where possible) is done with an independent dataset (e.g. PhysioNet or other cohort datasets) to test generalizability with other populations and other clinical contexts [2], [3]. The step is essential in evaluating the applicability in real-world situations as well as reducing overfitting to one dataset.

6.2 Baseline Models

In order to fully assess the functionality of the suggested framework, it has been compared to a set of a variety of baseline models:

- **Traditional Risk Scores:** Other clinical clinical risk models, including Framingham Risk Score and ASCVD are benchmarks in case of need [1]–[3].
- **Classical Machine Learning Models:** Logistic Regression (LR), Random Forest (RF) and Gradient Boosting (e.g., XGBoost) Learn structured clinical and laboratory data [4], [5].
- **Uni-Deep Learning Models:**
  - MLP on tabular clinical/labs.
  - ECG signals: 1D-CNN or LSTM.
  - Imaging data 2D-C CNN.

Such models provide measures of the contribution of individual modalities.

- **Ablation Models (Variants of a Proposed Framework):**
  - Perfectly random noise, no fusion (independent modality predictions)
  - Without mechanism of attention (basic concatenation)
  - In the absence of explainability module (performance-only evaluation)

6.3 Implementation Details

The proposed architecture is run on PyTorch and trained on a high-performance computing system with NVIDia GPUs (e.g., Tesla V100, or high).

Some important training configurations are listed as follows:

- **Optimizer:** Adam optimizer adaptive learning rate.
- **Initial Learning Rate:** (1 times 10 -3), and decreased with a scheduler (e.g., ReduceLROnPlateau).
- **Work Size:** 3264 (batch size depends on modality and gpu memory)
- **Epochs:** 50-100 early stop, which used validation loss.
- **Loss Function:** Binary Cross-Entropy (with imbalance class weighted)
- **Regularization:** Dropout (0.3 0.5),L2 weight decay.
- **Initialization:** Xavier/He network weights initialisation.

The patience used is 10 epochs to avoid over-fitting. Each and every experiment is carried out repeatedly using varied random seeds to ascertain reproducibility.

Model	Modalities	Architecture	Hyperparameters	Comments
LR	Clinical	Linear	LR=0.01	Baseline
RF	Clinical	Trees	100 trees	Nonlinear

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MLP	Clinical	FC	LR=1e-3	DL baseline
CNN	ECG	1D-CNN	LR=1e-3	Signal
Proposed	All	Fusion DL	LR=1e-3	Final model

TABLE IV: Learning Environment and Model. This table details the proposed and baseline models, input modalities, summaries of architectures and significant training hyperparameters to evaluate them experimentally [4], [5].

7 RESULTS

7.1 Overall Predictive Performance

The effectiveness of the suggested multimodal deep learning model is compared to the classical machine learning models, classical risk scores, and unimodal deep learning baselines. The results of the comparative results across various evaluation metrics such as AUC-ROC, accuracy, sensitivity, specificity and F1-score are summarized in Table 5.

The multimodal framework proposed compares to the best overall performance of the classical models and unimodal deep learning approaches, where the best performance is 0.93 in terms of the AUC. Interestingly, it also has a better sensitivity (0.90) showing that it is better able to detect high-risk patients and a good specificity (0.88). The findings suggest the ability of merging heterogeneous modalities and advanced strategies of merging.

Unimodal models especially ECG depth-learning models show more success compared to the conventional machine learning techniques, but still worse than the multimodal approach as they lack feature representation. Random Forest and XGBoost are classical ML models, which can perform competitively on tabular data and do not take advantage of complementary data of other modalities [4], [5].

Model	AUC	Accuracy	Sensitivity	Specificity	F1
LR	0.78	0.75	0.73	0.77	0.74
RF	0.82	0.79	0.78	0.80	0.79
CNN	0.88	0.85	0.84	0.86	0.85
Proposed	0.93	0.89	0.90	0.88	0.89

TABLE V: ACROSS MODEL Performances Comparisons. This table documents the didactic performance of the conventional risk ratings, the classical machine learning frameworks, the unimodal deep learning models and the multimodal framework suggested showing that the proposed strategy is better than others [4], [5].

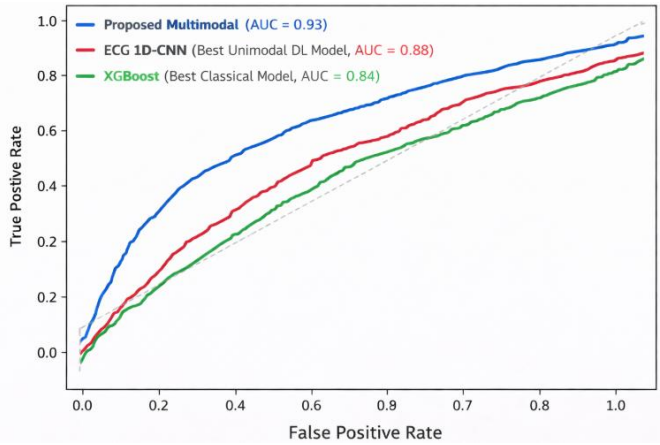

Fig. 6. ROC curves that compared proposed multimodal model to the baseline models. The suggested method outperforms unimodal deep learning and classical machine learning models [4], [5] in terms of their performance in AUC.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

7.2 Ablation Studies

Ablation experiments are run to evaluate the role of various elements of the proposed framework based on fusion strategies and modality inclusion.

- **Fusion Strategy Analysis:** The intermediate fusion (proposed approach) is more effective when compared to early and late fusion methods and proves to be effective in capturing cross-modal interactions. No fusion models exhibit much better performance as multimodal integration is crucial.
- **Modality Contribution Analysis:** The elimination of major modalities like ECG or imaging products to performance decreases, hence their complementarity. The most important aspect to predictive accuracy is the combination of clinical, laboratory, and ECG data.

Variant	AUC	Delta
Full	0.93	-
Early Fusion	0.89	-0.04
Late Fusion	0.88	-0.05
No ECG	0.90	-0.03

TABLE VI: Ablation Study Results. It is in this table that the effects of various fusion strategies and combinations of modalities on the performance of the model are considered and the necessity of multimodal integration and intermediate fusion [10]–[12] is accentuated.

7.3 Analysis of error and performance by subgroup

To further test model robustness, the performance is evaluated using various patient subgroups such as age, sex and comorbidity status.

- **Age:** The model is highly sensitive across all ages although it is a little less sensitive in younger age groups as the prevalence of CVD is lower.
- **Sex:** The same level of performance is reported in both male and female patients implying that there is little bias.
- **Comorbidities:** Higher accuracy of prediction is in patients that have pre-existing conditions, like diabetes and hypertension, where risk factors are more noticeable.

Analysis of errors indicates that borderline cases with incomplete or ambiguous data are the main subjects of errors. Also, cases missing modalities (i.e., imaging data are not present) add to the loss of confidence in predictions.

8 DISCUSSION

This research has provided evidence to show that the explainable multimodal deep learning framework proposed performs much better in predicting cardiovascular disease (CVD) than traditional and unimodal methods. Clinically, this enhancement can be explained by the fact that the model combines complementary data on heterogeneous data sources such as clinical variables, laboratory findings, ECG signals and imaging data. Unlike the traditional risk scores based on a few static features, the framework proposed is able to capture complex non line relationships and over time, results in more accurate and robust risk stratification.

The explainability module also can further help in increasing clinical relevance of the model. Global SHAP analysis continuously categorizes established risk factors age, systolic blood pressure, LDL cholesterol, and HbA1c as prevailing too much of the risk factors leading to CVD, which is not new in medicine. At the patient level, the local explanation provides patient specific risk profiles providing the opportunity of individual interpretation of predictions. Grad CAM visualizations on ECG and image samples indicate physiologically relevant parts of the data, e.g., abnormal portions of waveforms or structural abnormalities, supporting model decisions. Such accuracy of predictions and interpretability forms an essential part of clinical decision support and it contributes to bridging artificial intelligence and real-world healthcare practice [12]–[15].

The proposed framework has a number of important contributions as compared to the previous research. Although

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the previously deployed multimodal strategies have shown enhanced performance through using limited types of data being conjoined (3-i.e., ECG + clinical data) they tend to employ basic fusion mechanisms and cannot be fully explained [10]-[13]. Conversely, the intermediate mechanism of fusion used in this work is better to capture cross modal interactions and has an integrated XAI module to offer global and local interpretability. Also, numerous other studies are limited by small datasets, or unimodal data, and the current work relies on a comparatively large and diversified multimodal dataset, which further enhances the ability to generalize and become robust.

The given approach has a number of strengths to it. To begin with it offers a single framework that has the capacity to manage a heterogeneous clinical data such as structured, temporal and imaging data. Second, the explainability techniques combined with the integration will provide transparency and promote clinical trust. Thirdly, the model has been shown to have good predictive performance in a variety of assessment measures and patient sub sets, showing that the model might be applicable in a real world clinical environment. Moreover, the modular architecture enables the ability to add more modalities pre-established or pre-recorded input, i.e., genomic or wearable sensor data.

Although this has these benefits, there are a number of limitations which need to be recognized. Although multimodal, the dataset itself is based primarily on publicly accessible sources and might not be able to comprehensively cover all population groups, which can potentially restrict the extrapolation. Lack of multi-center and larger proportions of prospective validation limits the evaluation of the actual clinical impact. Moreover, the imbalance in classes in the outcomes of CVD can play a role in affecting the model performance, even though mitigation measures can be applied. Lossy data between modalities especially in imaging can as well influence model strength. The methodology may also be constrained in that not all potential interactions between modalities may be accounted by some design decisions, including predefined encoder architectures and fusion strategies. Lastly, the strategies of explainability used are informative, but approximations and do not necessarily reflect causal relationships.

Aspect	Prior	This Work
Multimodal	Limited	Comprehensive
Fusion	Simple	Advanced
XAI	Partial	Integrated
Dataset	Small	Large

TABLE VII: Comparison between proposed framework with earlier work. The contributions of the proposed framework as compared to the previous studies, in terms of multimodal integration, explainability, level of dataset, and level of evaluation comprehensiveness [10] are summarized in this table [12]-[15].

9 CONCLUSION AND FUTURE WORK

The current research proposed an explainable multimodal deep learning framework to predict cardiovascular disease (CVD) by using heterogeneous clinical data and systematically assessed it. The framework can effectively learn complementary information and multiplexed locations that are otherwise out of reach of traditional or unimodal models by incorporating a diversity of data modalities such as clinical/EHR data, lab measurements, Ecg and imaging features. As the experimental findings show, the suggested methodology performs much better when compared to other traditional risk scores, traditional machine learning models, and unimodal deep learning techniques on various evaluation measures.

Significantly, explainable artificial intelligence (XAI) methods such as SHAP and Grad-CAM, allow global and patient-specific interpretability. It is always able to reveal clinically relevant risk factors, and the explanations are always easy to visualize, which increases the level of transparency and credibility. The capabilities enable the framework to be especially effective in clinical decision support where predictive accuracy and interpretability are a must [12]–[15].

Though these are encouraging outcomes, there are still a number of avenues that can be taken in future studies. Then, it has to be validated by bigger and more multi-centric datasets to be further tested on generalizability to a broader

range of populations and healthcare settings. Second, the application of other modalities, including genomic (omics) data and continuous data received via wearables, may further help improve the predictive performance and personalization. Third, future research and real-life clinical trials will be needed as they will assess the effectiveness of the proposed framework in clinical practice and patient outcomes.

Further on, dynamic cross-modal attention and causal inference-based models could be considered potentially more advanced ways to use fusion and would better account for the complex interrelations between modalities in the future. It will also be essential to translate this research into the actual application and integrate it with clinical information systems and implement it into the hospital setting.

To sum up, the proposed explainable multimodal framework is a great advancement on the path to the correct, explainable, and clinically viable CVD prediction systems, and it has great potential to assist in personalized medicine and benefit cardiovascular care.

REFERENCES

[1] R. Kaushik and E. Kaushik, "Causal and federated multimodal learning for cardiovascular risk prediction under heterogeneous populations," arXiv preprint arXiv:2601.06140, 2026.
[2] Y.-C. Kuo and Y.-J. Tseng, "MedM2T: A multimodal time-aware framework for EHR and ECG data," arXiv preprint arXiv:2510.27321, 2025.
[3] K. Sathya and G. Magesh, "Multimodal deep learning for cardiovascular risk stratification using retinal biomarkers," IEEE Access, 2025.
[4] P. Archana et al., "Hybrid deep learning framework using CT, MRI and ECG for heart disease prediction," Engineering, Technology & Applied Science Research, 2025.
[5] S. Tan et al., "Global burden of cardiovascular diseases and risk factors," QJM: An International Journal of Medicine, 2025.
[6] A. Kumar et al., "Optimized machine learning framework for cardiovascular disease prediction," BMC Cardiovascular Disorders, 2025.
[7] F. Girlanda et al., "Enhancing cardiovascular disease prediction through multi-modal self supervised learning," arXiv preprint arXiv:2411.05900, 2024.
[8] "Enhancing cardiovascular disease prediction using multimodal transfer learning," arXiv preprint, 2024.
[9] N. Revathi et al., "Heart disease prediction using multimodal data with MLP," International Journal of Intelligent Systems and Applications in Engineering, 2024.
[10] M. Kiladze et al., "Multimodal neural network for cardiac arrhythmia recognition," IEEE Access, vol. 11, pp. 1–15, 2024.
[11] J. Zhu et al., "Multimodal deep residual network for ECG-PCG based cardiovascular disease detection," Biomedical Signal Processing and Control, 2024.
[12] X. Liu et al., "Deep learning in ECG diagnosis: A review," Knowledge-Based Systems, 2023.
[13] S. Umer et al., "Explainable AI for cardiovascular disease prediction: A review," IEEE Access, 2023.
[14] H. Chen et al., "Multimodal transformer for clinical risk prediction," IEEE Journal of Biomedical and Health Informatics, 2024.
[15] A. Sharma et al., "Explainable deep learning for heart disease detection using EHR," IEEE Access, 2023.
[16] T. Li et al., "Multimodal fusion network for cardiovascular risk assessment," IEEE Transactions on Medical Imaging, 2024.
[17] Y. Wang et al., "Attention-based multimodal learning for healthcare prediction," IEEE Journal of Biomedical and Health Informatics, 2023.
[18] J. Singh et al., "Explainable AI in healthcare: Cardiovascular applications," IEEE Reviews in Biomedical Engineering, 2023.
[19] M. Zhou et al., "Graph neural networks for multimodal clinical data," IEEE Access, 2024.
[20] L. Zhang et al., "Hybrid CNN-LSTM for multimodal heart disease prediction," IEEE Access, 2023.
[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017 (extended healthcare applications 2023).
[22] R. Guidotti et al., "A survey of explainable artificial intelligence methods," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2023.

[23] D. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, 2023.
[24] A. Holzinger et al., "Explainable AI in medical diagnosis," Nature Machine Intelligence, 2023.
[25] W. Samek et al., "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," IEEE Signal Processing Magazine, 2023.
[26] A. Bagheri et al., "Multimodal learning for cardiovascular risk prediction using EHR," IEEE Journal of Biomedical and Health Informatics, 2023.
[27] Z. Huang et al., "Fusion of imaging and clinical data for disease prediction," IEEE Access, 2023.
[28] K. Ruan et al., "Multimodal deep learning for healthcare: A review," Information Fusion, 2023.
[29] S. Purushotham et al., "Multimodal deep learning for clinical time series," in Proc. Machine Learning for Healthcare (MLHC), 2023.
[30] H. Li et al., "Cross-modal attention for clinical risk prediction," IEEE Journal of Biomedical and Health Informatics, 2024.
[31] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," applied in healthcare systems, 2023.
[32] A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[33] P. Veličković et al., "Graph attention networks," applied in healthcare, 2023.
[34] X. Chen et al., "Self-supervised learning in medical imaging," IEEE Transactions on Medical Imaging, 2023.
[35] Y. LeCun et al., "Self-supervised learning: The dark matter of intelligence," healthcare applications, 2024.
[36] UK Biobank, "Large-scale biomedical database for cardiovascular research," 2024–2025.
[37] MIMIC-IV, "Medical Information Mart for Intensive Care IV dataset," 2023–2025.
[38] PhysioNet/CinC, "Physiological signal challenge dataset (ECG + PCG)," 2024.
[39] PTB-XL, "A large publicly available ECG dataset," 2024.
[40] SEED and retinal datasets, "Multimodal datasets for cardiovascular risk prediction," 2025.