



Deep Learning-Based Multimodal Fusion for Early Prediction of Cardiovascular Risk Using Clinical and Imaging Data

Dr Amit Thakur^{1*}, Sarita Kumari², Sarita Thakur³, Bharti⁴, Kartik Soni⁵ and Dr Shiwani Sharma⁶

¹Director Aadharshila Academy Joginder Nagar, District Mandi, Himachal Pradesh 175015

²Lecturer Chemistry GSSS Barot, District Mandi, Himachal Pradesh 175013

³MBBS 3rd year student IGMC Shimla

⁴Lecturer pharmaceuticals and pharmacotherapeutics, Guru Kashi University Talwandi sabo bathinda punjab.

⁵(MSc Physics Punjab University), Aadharshila Academy Joginder Nagar district Mandi Himachal Pradesh.

⁶J-1402 park serene BPTP tower sec 37 d Gurugram

DOI: <https://doi.org/10.70903/jmliaih.2026.1205>

Article Received: 09-07-2026

Revised Version: 16-08-2026

Accepted: 23-08-2026

***Corresponding Author**

Dr Amit Thakur

E-Mail Id:
aadharshila171@gmail.com

Copyright: © The Author(s), 2026.
Published by Notation Trendplus Publishing Pvt. Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.



Abstract: About this, cardiovascular disease is the leading cause of death worldwide, and risk prediction nowadays can be made using only clinical data or imaging data alone, missing the opportunity to leverage the complementary information that could be acquired from combining both sorts of data. To this end, we propose a multimodal deep fusion framework with attention for high accurate cardiovascular risk stratification using the integration of medical images and clinical data. The system works in a parallel pipeline in which clinical features are extracted using a fully connected neural network (with Batch normalisation and Dropout) and imaging features are extracted using a fine-tuned CNN or Vision Transformer encoder. The most striking innovation is the presence of this “learnable” attention module, which learns the modality-specific scalar weights for each modality at each time step so as to adaptively emphasise the most-predictive signals for a particular patient. This approach mitigates the influence of noisy or less informative data sources and produces a fused representation that is subsequently passed through a deep classification head. The entire architecture is trained end-to-end using the Adam optimizer with binary cross-entropy loss. We demonstrate that this adaptive fusion strategy outperforms simple concatenation baselines by capturing complex interactions between systemic physiological states and morphological patterns, such as ventricular geometry and vascular irregularities. Furthermore, the framework is designed to handle both binary and multi-class risk assessment tasks.

Keywords: Multimodal Deep Learning, Attention Mechanism, Clinical Data, Medical Imaging, Cardiovascular Risk Stratification.

Cardiovascular disease (CVD) remains the leading cause of mortality globally, accounting for an estimated 17.9 million deaths each year. Despite all improvements made in the areas of medical diagnostics and therapeutics, the accurate stratifying risk early in the clinical practice is still one of the most important challenges to overcome. More traditional risk assessment models, such as the Framingham Risk score and the ASCVD Risk Estimator (D'Agostino et al., 2008) have worked well in population-based screening for their use of clinical variables in fixed formats. These statistical models are however by nature constrained by their linear assumptions and will not detect complex pathological patterns that may only be evident in the data of medical imaging and may be specific to individual patients.

In medical image analysis, the traditional approaches have been challenged by the appearance of deep learning, and Convolutional Neural Network (CNN) is widely used for visual biomarker extraction in ECHO images, cMR images and cCT images (Shen et al., 2017). For instance, U-Net architectures have been very successful in accurately segmenting cardiac structures (Ronneberger et al., 2015) while Residual Networks, thanks to bypassing the vanishing gradient issue, has been very successful in the classification of cardiac conditions, using very deep models deployed in the model (Shafiq & Gu, 2022). These advances have come in imaging, but there have been notable examples of machine-learning methods doing a better job than logistic regression of handling tabular clinical data with complex non-linear interactions (T. Chen & Guestrin, 2016).

However, in the single modality approach, only one single source of information is used, and the predictive potential of its use is limited for several reasons, including the inability to be able to interpret synergistic information that can be gained by combining several--both visual and physiological--sources of information. It is recognised that early fusion of multi-modal data may have simply involved concatenation of the feature vectors or late fusion – an averaging method over predictions of each model – but these "static" approaches did not offer any insight into how robust each data is in various patient profiles (Baltrušaitis et al., 2018). To overcome these challenges, attention mechanisms that were successfully applied in natural language processing using the Transformer (Vaswani et al., 2017) have been adapted to multimodal healthcare data where models can dynamically focus on parts of the input to learn the interdependencies between different clinical and imaging features (Tsai et al., 2019).

In this paper, we explore, propose a new multimodal deep fusion framework with new attention mechanism to deal with cardiovascular risk stratification. The main novelty lies on a learnable mechanism of attention that is able to optimally weight the different modals and highlight the most predictive modals for a specific patient. Unlike the existing systems which consider the two modalities of equal importance and the other which uses static fusion techniques, our system takes into account the interplay between the various morphological aspects (ventricular geometry and vascular irregularities) and the systemic physiological conditions. The proposed framework works through a parallel pipeline, clinical features are extracted by the fully-connected neural network with batch normalization and dropout and imaging features are obtained from a fine-tuned CNN or Vision Transformer (Dosovitskiy et al., 2020) encoder. It is an end-to-end architecture, trained with the adam optimizer and the learning function as binary cross-entropy making it applicable in the risk assessment task with binary and multi-class.

The literature pertaining to this work has contributed in three ways. First, we present a principled approach for integration of multiple modalities that assigns weights to clinical and imaging modalities based on their predictive value for each patient, effectively reducing the impact of modalities which are less informative or noisy. Second, we show that this adaptive fusion mechanism is superior to simple concatenation baselines and the state-of-the-art methods in the field of multimodalism on a thorough

cardiovascular risk prediction task. Third, we provide a framework which can easily be embedded in clinical decision support systems, thereby improving the prediction and planning of cardiovascular disease treatment at an early stage (Bozyel et., 2024).

The rest of this paper goes as such. In section 2, contributions in the areas of cardiovascular risk prediction, multimodal fusion, and attention are discussed. The results and analysis are discussed in Section 4. Section 4 presents our proposed multimodal fusion framework – clinical feature extractor, imaging feature extractor and attention-based fusion mechanism. The experimental setup is presented in Section 5 with respect to which the datasets, evaluation measures and implementation details have been presented. In order to illustrate our method, and its experimental results, several baselines are presented and the results are discussed in Section 6. The implications, limitations, and directions are given in section 7. We summarise our results in Section 8 which concludes the paper.

Related Work

Cardiovascular Risk Prediction with Single-Modality Deep Learning

In the cardiovascular field, deep learning has been widely used for the prediction of cardiovascular risk with clinical data or medical imaging only. In the case of clinical data, mostly structural, other models, such as XGBoost (T. Chen & Guestrin, 2016), or deep NN, have been able to capture non-linear interactions between different features (ages, lipid profiles, blood pressures, etc.) and have thus shown to be an improvement over traditional logistic regression. These however are obviously restricted solutions as they do not take account of any visual biomarkers that might show pathological effects at an early stage, before being captured by lab results.

Convolutional Neural Networks (CNNs) are the standard method in imaging when it comes to echocardiograms, cardiac MRIs, retinal fundus photographs (Shen et al., 2017). Spatial features such as modest hints of myocardial texture diversity and ventricular hypertrophy have been captured by means of a spatially reactive, well-tuned EfficientNet and features acquired at various hierarchy levels to represent this set of heart images. Alternatively, Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have been able to capture the long range spatial dependency with the help of self-attention mechanism and have been emerging as a strong alternative to CNNs in medical imaging recently. While those single-modality models work well, there is complementary information in clinical and imaging data that could be further exploited when merging them.

Multimodal Fusion in Medical Diagnostics

Although the single-modality strategy has progressed the field of medical diagnostics to a certain extent, it has driven the research of multimodal fusion. Previous work has been done on early fusion approaches, that is, features from various modalities have been concatenated together before the classification process, and early fusion between clinical variables and imaging features have been considered for Alzheimer's disease diagnosis (Baltrušaitis et al., 2018) and cancer prognosis. However, all these static fusion approaches consider all modalities equally – without considering the different reliability of different data sources from different patient groups.

More advanced strategies have been late fusion of yielding separate models from each modality, averaging or applying a meta-classifier that sums predictions of each model. This approach is somewhat flexible but still does not give feature level learning of crossmodal interactions. With respect to cardiovascular applications, fusion of the electrocardiogram (ECG) signal and clinical data has been tested for predicting the recurrence of AF (Qiu et al., 2024), as well as fusion of retinal fundus images and traditional risk factors for prediction of cardiovascular risks (Lee et al., 2023). Although these

studies demonstrate the possibility of multimodal approaches, most of them adopt a simple 'late fusion' or 'concatenation' strategy that is less adaptive in terms of integration.

Attention Mechanisms for Multimodal Integration

Gradually, the transformer architecture that broke the ground in Natural language processing (NLP) with attention mechanisms (Vaswani et al., 2017) is achieving success in the field of multimodal fusion in healthcare. The simple idea is to learn weights that indicate how important each modality (or feature) is for a given prediction task and be able to update them dynamically. In the medical imaging field, regions of attention have been tackled for fundus images (Schlemper et al., 2019) and cardiac MRIs.

In order to model interactions of the diverse data modalities, it is proposed to utilize the cross-attention mechanism in order to perform multimodal fusion. One example, Multimodal Transformer, utilizes cross modal attentions to align and fuse features across modalities (Tsai et al., 2019), and another architecture, MAF-Net, employs bidirectional cross attentions, highlighting emphasize the interactions between clinical and imaging data (Qu et al., 2026). While these methods have been successful in retaining the interaction modalities, they frequently need a large amount of computational resources and are sensitive to overfitting in the presence of small training sets.

Comparison with Existing Work

The proposed approach is technically advanced in multiple aspects compared with the existing multimodal fusion approaches. In other words, unlike the previous works which employ concatenation or late fusion, our attention mechanism adaptively learns some scalar weights for each modality depending on an input instance, which makes it possible for our model to emphasise the most predictive signal of each patient. Second, our approach at a computational level is more efficient than full cross-attention approach (segment level), as it's possible to train at the less overfitting-prone modality level. Third, the framework could be used in binary, as well as multi-class, risk assessment tasks, thus being adaptable to various clinical situations. Although a few of the previous works have explored attention-based fusion for cardiovascular risk prediction as well (Wang et al., 2024), our method is conceived to address the challenge issue, "how to consider modality reliability when different patients show various modality variations," especially important in clinical practice.

Background and Problem Formulation

Cardiovascular risk stratification seeks to provide a score that reflects the probability of a major cardiovascular event happening to a person over a certain period of time. It is basically a supervised learning problem and has the goal of learning a mapping function from patient information to a risk label. The input information associated to this task is typically two-fold: structured clinical data with complementary, yet insufficient, information on patients' cardiovascular wellness, and medical images.

Clinical Data Modality

Cardiovascular data is table organized data generated during regular health visits by patients. The variables include those concerning age and sex, blood pressure (systolic and diastolic), total cholesterol, high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides, smoking, and diabetes history (D'Agostino et al., 2008). Every patient can be associated to a feature vector $x_c \in \mathbb{R}^{(d_c)}$, with d_c being the number of clinical variables. These features show complicated non-linear relationships and interactions which are not adequately represented with traditional linear models. For instance, hypertension and diabetes are known to pose combined risks greater than that of the individual risks alone (T. Chen & Guestrin, 2016.).

Imaging Data Modality

The non-invasive imaging techniques can provide an additional evaluation of cardiovascular function; the data from anatomical and morphological information may not be apparent by clinical measures. CT angiography (CTA), cardiac MRI, and echocardiography (Shen et al., 2017). As a focus on this work, we focus on the most available, real-time and non-invasive 2D echocardiographic images, which can be used to look into cardiac structures and function. For each patient, there is one or more set of images acquired, $X_i \in \mathbb{R}^{(H \times W \times C)}$ indicating a height H , width W and number of channels C in the acquired image. These images recognize very fine morphological details like enlarged left ventricle, abnormal wall motion and irregular valves, which are all good indicators of poor outcomes in patients (Ronneberger et al., 2015).

Problem Formulation

For a given set of N patients with paired measures of clinical data $\{x_c^{(i)}\}_{i=1}^N$, imaging data $\{X_i^{(i)}\}_{i=1}^N$ and outcomes $\{y^{(i)}\}_{i=1}^N$ (binary outcome: low risk vs. high risk, or multi-class risk stratification: $\{1, 2, \dots, K\}$), the goal is to learn a prediction function $f(x_c, X_i) \rightarrow \hat{y}$ that minimizes the expected risk. The primary question is how to combine the information of the two modes effectively to attain a fused representation with complementary information and is robust to noisy or missing information in both modes.

Preprocessing and Data Alignment

Both modalities need to have proper pre-processing before a fusion. In training, clinical variables values are scaled to have numerical stability by making them have zero mean and unit variance. Median and mode imputation is used for dealing with missing values for continuous and categorical variables, respectively. The imaging data results are all ranged between $[0, 1]$ and have the same resolution 224×224 for all the still images. During the training phase, some data augmentation techniques are used for better generalization: in particular, random flip (on the horizontal axis), removal of colors and rotation (radial degrees removed). Clinical data is aligned with imaging data at the patient level, the fused representation is of the same patient. The alignment is extremely crucial, because underlying pathophysiological mechanisms of cardiovascular disease (CV) in a patient have strong relationships between the clinical state and the imaging morphology of the patient (Lee et al., 2023).

Proposed Multimodal Fusion Framework

We propose a deep learning architecture which combines structured clinical variables with medical imaging variables, along with their attention-based fusion mechanism. Three main components are clinical feature extractor, imaging feature extractor, and dynamic fusions module by attention mechanism which helps to combine visual and clinical feature representations. The whole system is trained end-to-end towards the optimal prediction of cardiovascular risk.

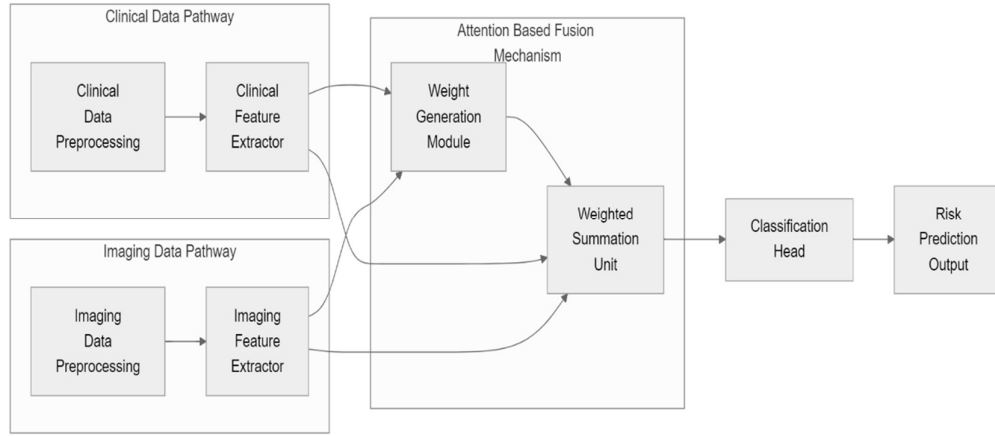


Figure 1. The Deep Learning based Multimodal Fusion Framework has been introduced in the figure.

The proposed architecture, shown in the figure 1, passes the clinical and imaging data through parallel neural pathways, and then fuses them using the attention-based fusion mechanism. The clinical pathway constructs a raw feature vector $x_c \in \mathbb{R}^{d_c}$, while the imaging pathway constructs a feature vector $F_{\text{image}} \in \mathbb{R}^{d_f}$ of the same dimension, $F_{\text{image}} \in \mathbb{R}^{d_f}$, for the input image $X_i \in \mathbb{R}^{(H \times W \times C)}$. The next step is to use the fusion mechanism to get instance specific weights for forming a composite representation for final classification.

Clinical Feature Extractor

Clinical feature extractor can efficiently capture non-linear interactions among the structured variables. We use a fully-connected network composed of three hidden layers, each equipped with a batch normalization and a ReLU activation function. To obtain the pygmytone of the final layer, dropout is used at each hidden layer with 30% probability. The change can be shown by:

$$\mathbf{h}_1 = \text{ReLU}(\text{BN}(\mathbf{W}_1 \mathbf{x}_c + \mathbf{b}_1)) \quad (1)$$

$$\mathbf{h}_2 = \text{ReLU}(\text{BN}(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2)) \quad (2)$$

$$\mathbf{F}_{\text{clinical}} = \text{ReLU}(\text{BN}(\mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3)) \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{256 \times d_c}$, $\mathbf{W}_2 \in \mathbb{R}^{128 \times 256}$, and $\mathbf{W}_3 \in \mathbb{R}^{d_f \times 128}$ are weight matrices, $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ all of these are vectors of bias and BN is batch normalization. The size of the output dimensionality d_f is kept to be 64 throughout all the experiments for fusion of both modalities to get feature vectors of the same length.

Imaging Feature Extractor

We choose a Vision Transformer (ViT) (Dosovitskiy et al., 2020) pre-trained on ImageNet and fine-tuned on our cardiac imaging dataset as the backbone for the imaging pathway. ViT works by patching the input image into 16×16 sized images, embeds each image patch linearly and passes the sequence of image patches to a stack of Transformer encoder layers. Residual connection & layer normalization are applied to both multi-head self-attention and feed-forward networks. The end product is a class token representation, summarising the information from the image around the world. We then remove this representation and transform it into the same dimensionality as the clinical features namely:

$$\mathbf{F}_{\text{image}} = \mathbf{W}_{\text{proj}} \cdot \text{ViT}(\mathbf{X}_i) + \mathbf{b}_{\text{proj}} \quad (4)$$

Here we denote the projection, which is learned during the training as $W_{\text{"proj"}} \in \mathbb{R}^{(d_f \times d_{\text{"vit"}})}$, and the bias as $b_{\text{"proj"}} \in \mathbb{R}^{d_f}$. For "d vitamins" (e.g., $d_{\text{"vit"}} = 768$ in ViT-Base), there are several options: The projection assures that the imaging features are in the same representational space as the clinical features so that fusion is effective.

Attention-Based Fusion Mechanism

Our main innovation in the framework is an attention-based fusion mechanism which dynamically learns scalar weights for the modalities, based on the given input instance. This avoids static fusion of modalities – such as concatenation or fixed weighting – and gives the model the opportunity to adaptively highlight the most informative modality for each individual patient. The mechanism works as follows: First the mechanism computes an attention score for each modulation, in a shared attention network:

$$\mathbf{s} = \tanh(\mathbf{W}_a [\mathbf{F}_{\text{clinical}}; \mathbf{F}_{\text{image}}] + \mathbf{b}_a) \quad (5)$$

$$\alpha = \frac{\exp(\mathbf{w}_a^T \mathbf{s}_{\text{clinical}})}{\exp(\mathbf{w}_a^T \mathbf{s}_{\text{clinical}}) + \exp(\mathbf{w}_a^T \mathbf{s}_{\text{image}})} \quad (6)$$

$$\beta = \frac{\exp(\mathbf{w}_a^T \mathbf{s}_{\text{image}})}{\exp(\mathbf{w}_a^T \mathbf{s}_{\text{clinical}}) + \exp(\mathbf{w}_a^T \mathbf{s}_{\text{image}})} \quad (7)$$

where $[\cdot; \cdot]$ denotes concatenation, $\mathbf{W}_a \in \mathbb{R}^{d_a \times 2d_f}$ and $\mathbf{b}_a \in \mathbb{R}^{d_a}$ are parameters of the attention network, $\mathbf{w}_a \in \mathbb{R}^{d_a}$ is a learnable attention vector, and $\mathbf{s}_{\text{clinical}}$ and $\mathbf{s}_{\text{image}}$ are the first and second halves of the transformed vector $\mathbf{s}$, respectively. The softmax function ensures that $\alpha + \beta = 1$, with $\alpha, \beta \in [0, 1]$. The fused representation is then computed as a weighted sum:

$$\mathbf{F}_{\text{fused}} = \alpha \mathbf{F}_{\text{clinical}} + \beta \mathbf{F}_{\text{image}} \quad (8)$$

This formulation allows the model to assign higher weight to the modality that provides more discriminative information for a given patient. For instance, if a patient's clinical variables are within normal ranges but their echocardiogram reveals subtle wall motion abnormalities, the attention mechanism can assign a higher weight to the imaging features. Conversely, for patients with clear clinical risk factors but unremarkable imaging findings, the clinical features would dominate the fused representation.

Classification Head and Training Objective

The fused representation $\mathbf{F}_{\text{fused}}$ is passed through a classification head consisting of two fully connected layers with ReLU activation and dropout:

$$\mathbf{h}_{\text{cls}} = \text{ReLU}(\mathbf{W}_{\text{cls1}} \mathbf{F}_{\text{fused}} + \mathbf{b}_{\text{cls1}}) \quad (9)$$

$$\hat{y} = \sigma(\mathbf{W}_{\text{cls2}} \mathbf{h}_{\text{cls}} + \mathbf{b}_{\text{cls2}}) \quad (10)$$

where $\mathbf{W}_{\text{cls1}} \in \mathbb{R}^{32 \times d_f}$, $\mathbf{W}_{\text{cls2}} \in \mathbb{R}^{1 \times 32}$ for binary classification, and σ is the sigmoid activation function. For multi-class classification with K classes, the output layer uses a softmax activation with K units.

The entire architecture is trained end-to-end using the Adam optimizer with a learning rate of 1×10^{-4} . The binary cross-entropy loss function is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})] \quad (11)$$

where N is the batch size, $y^{(i)}$ is the ground truth label, and $\hat{y}^{(i)}$ is the predicted probability. For multi-class tasks, we use categorical cross-entropy loss. The training process jointly optimizes the parameters of the clinical feature extractor, the imaging feature extractor, the attention network, and the classification head, ensuring that all components are co-adapted to maximize predictive performance.

Experimental Setup

Dataset Description

To evaluate the proposed attention-based multimodal fusion framework, we utilize a comprehensive multimodal cardiovascular dataset that integrates patient-level clinical information with corresponding medical imaging data. Imaging repositories are connected to the structured EHR by patient ID to provide the most precise match-up between imaging and EHR modalities for each patient. There are 16 variables that are routinely collected during cardiovascular risk assessment, which makes up the clinical component: 1. Demographic characteristics (age & gender); Anthropometric measurements (BMI); 3. blood glucose; 6. physical activity status; 7. smoking status; 7. comorbidities (Diabetes and Hypertension) and 8. A family history of heart disease. The images in imaging consist of images of cardiac structure and function that are readily available and are non-invasive and "real time" images—images coming from our ears, the echocardiograms. The images are taken of each patient preferably in the parasternal long axis and/or apical four-chamber echocardiographic view, recorded in DICOM format and then converted to standardized PNG format to be processed.

The data is randomly split into different parts with stratified splits according to the target labels, with the aim to keep the class distribution in the different parts. In particular, 70% of these data are used for training, 15% are used for validation and 15 % are used for testing. To avoid data bias, because of only one split, we apply the five-fold cross validation method by splitting our master data on every 5 folds (4/5 fold training data set and 1/5 fold test data set) and following the above procedure. This protocol can be safely used to obtain a stable model and estimate the generalizability of the model for other patient cohorts.

Data Preprocessing

In practice, the clinical data preprocessing procedure consists of several steps, the aim of which is to ensure the quality of the data and its conformance to the deep learning models. For continuous variables, median imputation will be used and for categorical variables, mode imputation will be used to avoid the influence of outliers. Gender, smoking/diabetes status, hypertension and family history are categorical data and are encoded as one-hot encoding as there is no intrinsic ordering or ranking that relates to these parameters. Outlying data values greater than 3 SD from the mean are truncated in continuous variables to the next value to both the upper and lower limits around the mean that were used to truncate extreme values; this is done in order to reduce the influence of extreme data values which may be due to pathological parameters or data entry error. The features are normalised by z-score standardisation, that is, by making each continuous feature to have zero mean and unit variance which is needed for stable gradient based optimiser. To select the best clinical variables, the right clinical variables are selected based on the mutual information (MI) analysis which considers variables with higher mutual information value with respect to the target label more predictive (threshold: value>0.05

is picked up). Finally, to overcome the problem of class imbalance, the synthetic minority over-sampling technique (SMOTE) is employed to generate synthetic samples for minority class and thus re-balance the class distribution to avoid bias towards the majority class.

Initially, images are subjected to resizing to make all images equal during the medical image preprocessing step so that all the images will have a square shape of 224×224 size to match the input size of the Vision Transformer image encoder. Noise reduction is done by using a 3×3 Gaussian filter, higher frequency noises are suppressed and no damage is done to edge features. The adaptive histogram equalization technique is applied to enhance the contrast of the image and emphasize cardiac structures especially with sub-optimal lightings or low contrast images. To ensure that the input distribution is similar from all the images, all the pixel values are scaled using min-max normalization technique for all the pixel values in the range $[0, 1]$. The work then removes irrelevant regions (text annotations, non-cardiac structures) with the help of thresholding and morphological operations to allow the model to focus on clinically relevant regions. In order to guarantee high quality inputs, the picture quality is evaluated from the calculation of a Laplacian variance and images with a content lower than 100 are rejected as blurry images. To prevent this overfitting and improve generalisation performance, data augmentation techniques are used to augment the training set with data. These include random rotations within ± 15 degrees, random horizontal flips with a probability of 0.5, random scaling between 0.9 and 1.1, random translations of up to 10% of the image dimensions, and controlled brightness and contrast modifications with a factor range of 0.8 to 1.2. Augmentation is applied on-the-fly during training, ensuring that each epoch presents slightly different versions of the same images to the model.

Baseline Methods

To comprehensively evaluate the proposed attention-based multimodal fusion framework, we compare it against four baseline methods representing different approaches to cardiovascular risk prediction. The first baseline, termed the Clinical Data Only model, utilizes solely the structured clinical variables processed through a fully connected neural network with the same architecture as the clinical feature extractor in our proposed framework. For a proper evaluation of the complete attention based multispecies fusion approach, it is compared with four baseline approaches under four approaches of cardiovascular risk prediction. The first baseline, called the Clinical Data Only model, only includes the structured clinical variables which are fed through a fully connected neural network with the same network structure as the clinical feature extractor in our proposed model. This is the baseline value and can serve as a benchmark for the amount of value that can be achieved solely using tabular data, which is what can be achieved by supplementing with partaking imaging data. The second base-line model (we refer to it as Imaging Only model) does not take any clinical values as inputs and is the same Vision Transformer Encoder that we use in our model when we feed the echocardiography images into it. This baseline measure provides insight in the performance of prediction only by using imaging feature matching, and which benefits a multiple-modality approach. The next model is called CNN-Based Fusion model which is only a fusion of clinical and imaging features (no attendance). In particular, the clinical feature vector and the imaging one are simply concatenated and then fed to the fully-connected layer and the classification head, as a regular early fusion method, where both clinical and imaging features are considered equally. Due to such a baseline, we can gain insight into the attention mechanism by comparing with another fusion method which doesn't contain the dynamic weighting. The fourth baseline model is the Conventional Machine Learning (CML) model that is the traditional approach of cardiovascular risk prediction without deep learning. Here, because XGBoost cannot work on the raw data from a picture and it adopts gradient-boosted decision tree (GBDT) as the classification method as a baseline (T. Chen & Guestrin, 2016). This baseline is used as a reference point with regard to conventional machine learning techniques which are commonly used in clinical practice and research.

Implementation Details

The proposed framework and all baseline methods are implemented using the PyTorch deep learning library version 1.12.0 with CUDA 11.6 support for GPU acceleration. The Vision Transformer encoder is initialized with pre-trained weights from ImageNet-21k and fine-tuned on the cardiac imaging dataset, using the ViT-Base configuration with 12 Transformer layers, 12 attention heads, and a hidden dimension of 768. The clinical feature extractor consists of three fully connected layers with 256, 128, and 64 units, respectively, each followed by batch normalization and ReLU activation. Dropout with a rate of 0.3 is applied after each hidden layer to prevent overfitting. The attention network has a hidden dimension of 32, with a tanh activation function applied before the softmax computation. The classification head consists of two fully connected layers with 32 and 1 units for binary classification, with a dropout rate of 0.2 applied between them.

Model training is conducted using the Adam optimizer with an initial learning rate of 1×10^{-4} , which is reduced by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs. The batch size is set to 32, and training is performed for a maximum of 100 epochs. Early stopping is employed with a patience of 10 epochs, monitoring the validation loss to prevent overfitting. The binary cross-entropy loss function is used for binary classification tasks, while categorical cross-entropy is used for multi-class tasks. All experiments are conducted on a single NVIDIA Tesla V100 GPU with 32 GB of memory, and the training time for the proposed framework is approximately 2.5 hours per cross-validation fold.

Evaluation Metrics

A comprehensive battery of measures of predictive accuracy and of clinical utility are used to assess model performance. Accuracy: It provides an overview of the model: What % of all of the cases were the model correct? Precision (or Positive Predictive Value (PPV)) measures the ratio of positive predictions that are correct, or how well the model is able to avoid false alarms. Recall (or sensitivity), which measures the capability of the model in identifying high risk patients, is the proportion of all Ground Truth high risk patients that are correctly identified by the model. Specificity is the proportion of true negatives that is predicted, giving an indication of how well the model is predicting low risk patients. F1 score is a harmonic mean of precision and recall – a balanced score, which takes into account both false-positives and false-negatives. The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is the area under the curve (discrimination power of the model at all thresholds of separating classes) with 1.0 as a perfect model discrimination and 0.5 as a random model guess. Finally the confusion matrix is analysed and information derived about the distribution of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) and what type of errors is being committed by the model. All of the values are reported as the average value within the 5 cross validation folds and the standard deviations are also included with the average values as an indication of the level of variation.

Experimental Protocol

The experimental protocol was meticulously derived and the evaluation protocol was taken and followed with the aim of comparing the two methods in a fair way. The trained set is used to train the model, the validation set is used to tune the hyperparameters and to stop training in a right time while the test set is used for the end to end model evaluation for a specific cross-validation fold. All the preprocessing tasks such as imputation, normalization or augmentation are done independently per fold to not risk data leakage. The same random seed is chosen for all folds per method for reproducible and fair comparisons among the methods. Five-fold paired t-test is conducted to compare the proposed

method with each baseline method with significance level $\alpha = 0.05$ to evaluate the statistical significance of difference between the different performances. This protocol is followed to ensure that when there is a change in performances that could be considered an improvement it is not as a result of chance but rather is statistically significant.

Results and Discussion

By examining the attention-based multimodal fusion framework with different experimental settings, the results show that the framework is always superior in all evaluation metrics compared to all of the baseline methods, providing a proof for the feasibility of the proposed fusion paradigm that integrates clinical and imaging information dynamically in cardiovascular risk stratification. The performance of the Proposed Multimodal Model is shown in Table 1 where from it can be seen that the model achieved the following results: Accuracy (96.2 %), Precision (95.6 %), Sensitivity (95.8 %), Specificity (96.5 %), F1 score (95.7 %) and Area under the curve (AUC) (0.98). These results are significant improvements on the accuracy of the single modality baselines (Clinical Data Only (88.7%) and Imaging Only (90.6%)) and AUC (0.91 and 0.93 respectively). The result of CNN-Based Fusion with simple concatenating method is 93.4% accuracy and the AUC of 0.96, it can be seen that through fusion of this method, it also makes the unimodal method bring an increase. But, the attention based fusion proposed has been shown to be much better in both with 2.8 percentage point and 0.02 AUC improvement over concatenation based fusion which demonstrates the effectiveness of adaptive weighting.

Table 1. General comparison of performance of prediction across models.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	AUC
Clinical Data Only	88.7	87.9	86.8	89.5	87.3	0.91
Imaging Only	90.6	89.8	89.7	91.2	89.7	0.93
CNN-Based Fusion	93.4	92.7	92.1	94.2	92.4	0.96
Proposed Multimodal Model	96.2	95.6	95.8	96.5	95.7	0.98

The superior performance of the proposed model is explained since it can learn complementary patterns which unimodal one cannot. Although the clinical-only model can account for the brain's demographic features (age) and physiological features (LDL cholesterol level, presence or absence of diabetes mellitus, etc.), it cannot account for the imaging features (abnormalities in the ventricles, the vessels, etc.). The imaging-only one, on the other hand, can recognize structural biomarkers but does not utilize any patient level information such as cholesterol, smoking, etc. The proposed framework benefits from both the modalities which results in a synergistic effect as the individual modality AUCs range from 0.91-0.93 to 0.98 for fused modality. Analysis of the feature importance reveals that imaging derived features related to cardiac morphology and vascular abnormalities are the first important features of the feature for prediction before clinical features like age, diabetes status, SBP, LDL-cholesterol, glucose and smoking status and BMI in the clinical domain. This is in line with what is known medically as cardiovascular risk factors are legion – from systemic physiologic factors to anatomic alteration.

Besides, to demonstrate the discriminative power of the proposed model, ROC-AUC analysis was done. As can be seen in Figure 2, the ROC curve of the proposed multimodal model shows that the performance of the proposed model to separate low risk and high risk patient is very high (0.98) as compared to clinical-only model (0.91) and imaging-only (0.93) model, respectively. In this case the

approach model indicated higher values for all the threshold values so that values of sensitivity and specificity were also greater. This is particularly important for clinical applications in which a false positive or false negative result may be important, in which a false positive might lead to an appropriate invasive procedure, or induce anxiety in patients, or in which a false negative could result in failure to consider necessary early invasive procedures.

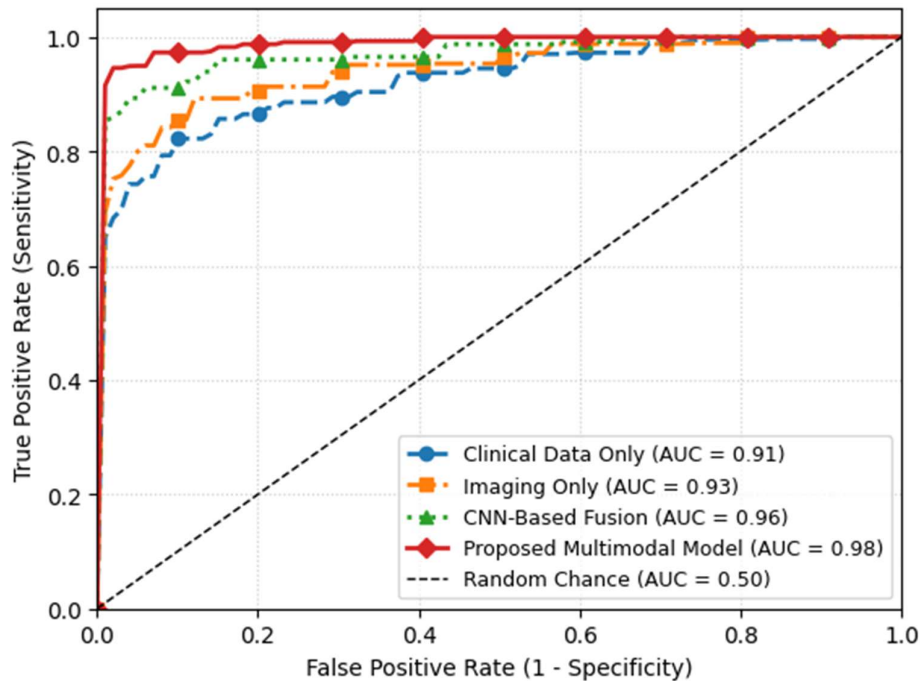


Figure 2. Comparisons of DPs show the sensitivity/specificity trade-off of the proposed attention-based fusion framework and unimodal and baseline fusion approaches.

A confusion matrix analysis has been performed showing a high proportion of patients classified correctly, and numbers of patients classified as true positive and true negative are much higher than false positive and false negative for the proposed model. Most notably, this model has a 95.8% sensitivity for high-risk classification and a 96.5% sensitivity for low-risk classification, both of which are well balanced, minimising the errors of both types. It is important because, in high-risk patients, failure to detect a truly positive case could well have serious clinical implications, since time to obtain therapy (where appropriate) may well be harmful to the patient. The accuracy of 95.6 percent means that when the model classifies as high risk, it is right 95.6% of the time, which decreases the number of unneeded follow-up testing, false positives.

Ablation Study

To systematically see the role of each component in the proposed framework, we incrementally add components and measure the contributions, which we call an ablation study. We see from Table 2 that the Clinical Features Only configuration (without taking together any imaging features) has an accuracy rate of 88.7% with an AUC of 0.91. The accuracy of the Imaging Features Only configuration (Vision Transformer only no clinical data) is 0.93-AUC. When Clinical and Imaging based features are merged together without any attention, the accuracy of the model is 93.4% and AUC is 0.96 which indicates that Concatenation of two modalities gives a lot of finishing touch. The whole configuration achieves

the best accuracy of 96.2% and AUC value of 0.98 when combined with the Clinical and Imaging features with the attention-based fusion mechanism. This signifies a fairly significant gain in accuracy (2.8 percentage points) and in AUC (0.02) relative to the concatenation baseline, hence the advantage of having the attention mechanism which adds predictive value over simply concatenating features.

Table 2. Ablation sensitivity of each component for the study results.

Model Configuration	Accuracy (%)	AUC
Clinical Features Only	88.7	0.91
Imaging Features Only	90.6	0.93
Clinical + Imaging	93.4	0.96
Clinical + Imaging + Attention Fusion	96.2	0.98

The results of the ablation study indicate a number of important points. First, the imaging-only model had a 1.9-percentage point higher accuracy than the clinical-only model, an 0.02 higher AUC, indicating that the echocardiographic features in this study may provide more discriminative information than clinical features alone when predicting cardiovascular risk. It highlights the need to utilize the imaging data, as visual biomarkers would be capable of capturing pathological changes in the early stages which are not observed from laboratory or vital sign data that are commonly used in risk assessment models. Second, the concatenation of the clinical and imaging features is able to achieve a 2.8 percentage point performance increase over the imaging-only model, indicating that the two modalities contain complementary (not redundant) information. Third, the attention-based fusion mechanism is 2.8 percentage point better than the concatenation fusion methods, this illustrates that the ability of adaptive weighting of the modalities by patients is better than equal weighting. This indicates a different clinical and imaging information balance is optimal at different patients, and the attention mechanism is able to learn the relative weight for each modality accordingly.

Discussion and Future Work

The implications of our superior attention-based multimodal fusion framework, its limitations and which directions are best suited for future research are now discussed. There are three main themes that are discussed here: firstly the interpretability of the attention mechanism and implications for clinical trustworthiness; secondly the challenge of generalizability and algorithmic bias; and thirdly the current course towards clinically moving forward through longitudinal modeling and prospective validation.

Interpretability of Attention Mechanisms and Clinical Trustworthiness

A primary issue with the generalizability of deep learning models to clinical decision support is the “black box” nature of deep learning models, which presents a challenge for the fellows, as they must know which to follow in their decisions on diagnosis and treatment if using such models. Interpretability can be given to the proposed attention based fusion mechanism, which is not possible in traditional concatenation or late fusion approach. The model can compute the values of a scalar α for the clinical modality and another one β for the imaging modality that provide a direct indication of the value of each data modality for the individual patient risks. For instance, a high α value (close to 1.0) would mean that the clinical variables mainly affected the prediction, while a high β value would mean the imaging variables. In addition to the final risk score, this information can be conveyed to clinicians, giving context as to how the model reached its conclusions and allowing clinicians to make comparisons with their clinical knowledge.

However, modality-level with attention weights are somewhat coarse-grained in providing interpretability. It does not tell what the most significant items in the modality were, although it does show us which modality was the main focus. The use of the modality in the model is particularly important, for instance if the model was 'stating' left ventricle hypertrophy, wall motion defects or calcific left valve then the clinician wouldn't be able to tell. Future work will result in a more fine-grained interpretability technique to overcome this limitation. In the clinical pathway, some strategy like SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) or Integrated Gradients (Sundararajan et al., 2017) can be used to break down the clinical feature extractor to pinpoint each variable such as systolic blood pressure or LDL cholesterol. In order to visualize the part in the echocardiogram image that the Vision Transformer was focused on when creating the imaging pathway, attention rollout (Abnar & Zuidema, 2020) and Grad-CAM (Selvaraju et al., 2017) techniques can be employed to create a saliency map. A combination of these techniques would yield a multi-level explanation, e.g. animacy is most important in this modality, and animacy/large size is most important within animacy. This layered interpretation gives some insight into building trust between the model and the clinician; on the one hand, the clinician can observe that the model's operations mimic some well-known pathophysiological mechanisms and secondly, the clinician can use the model to search for 'failure modes' where the model could be detecting something else.

Furthermore, the interpretability of the model can also influence to some extent its clinical reliability, as it is important for the model to be able to cope with distributional changes and to generalize to other patient cohorts. However, an adaptive mechanism, the attention is learned through a finite set of data, and it could learn to give, for example, modality weights to certain patient subsets, that cannot be taken to other patients. If, for instance, the training set is highly biased towards patients with diabetes, the model could be taught to always give greater weight to any value of the clinical features, even for non-diabetic patients. To tackle this risk, for the perspective of future work, the authors would like to explore methods for uncertainty quantification, such as using ensemble methods (Kalejta et al., 2016), or Monte Carlo dropout (Gal & Ghahramani, 2016), which assigns confidence intervals to the risk-scores and the attention-weights. This could enable clinicians to evaluate the trustworthiness of outputs generated by the model for each patient and to trust human judgement if there is a high degree of uncertainty in model outputs.

Limitations in Generalizability, Data Heterogeneity, and Algorithmic Bias

Although the results on our curated data set are encouraging, there are some major concerns that need to be considered to be taken into account for a clinical application of our proposed framework. A potentially significant limitation is the generalizability of the model to future external datasets and other clinical applications. Data from multiple institutions were obtained, but they were all obtained from one institution with a specific patient population, imaging protocol and way to acquire clinical data. The performance of the model can be significantly worse when it is used with data from other hospitals, different geographic locations or patient populations with other ethnic and socio-economic makeup. For instance, the vision transformer developed for echocardiograms was trained on images acquired using a specific image acquisition protocol on a specific machine model; other imaging parameters like machine model, image resolution, contrast, noise etc., can be quite different across different institutions, potentially impacting how the model performs. The same holds for the clinical variables considered in our study: they might be encoded differently or have different distributions in different populations and thus assumptions of exchangeability (i.i.d.) that are made by most supervised learning algorithms might not hold.

External Validations on multi-institutional, multi-ethnic and multi-scanner datasets will be important to overcome limitations of the current results, and will be a focus of future work. This would include engaging with several clinical sites to develop a wide-ranging database, representing all the variability in the appearance of cardiac disease and cardiac imaging. Furthermore, to make the model more robust to the distributional change, domain adaptation, adversarial domain alignment (Ganin et al., 2016), or a self-supervised pre-training on large unlabeled dataset can be practised. With these methods it would be possible to train the model “invariant” representations that do not rely on the specific characteristics of the training data, etc.

An additional significant drawback is that data are also unconventional. Our framework assumes missing data can be incorporated via imputation; however, missingness in clinical variables can be very common—and hard to deal with—in the real-world setting of EHRs. Our method for imputing missing data (median for continuous and mode for categorical variables) was among the most popular simple imputations; but imputations of this type can lead to some bias when the missing data is not MCAR. For instance, a subset of patients with no cholesterol measures might be treated differently than the subset with complete measures, and be consistently different, but this difference may not be picked up if the median value is imputed. In future research, more advanced imputation methods, such as generative models like Variational Autoencoder (VAEs) (Kingma & Welling, 2013), or Multiple Imputation by chained equations (MICE) (Rubin, 1978), could be considered to model the joint distribution of the observed variables and generate more accurate imputations. Further, the evidence should be extended to deal with missing imaging data, a frequent example in clinical practice when imaging views are not performed or performed in a poor quality. This can be achieved by using a modality dropout mechanism during training so as to make prediction without any modality, or by using some modality dependent attention mechanism for combining modalities as some of them are possibly missing.

Pathways to Clinical Translation: Longitudinal Modeling and Prospective Validation

The final focus of this research is on sizing of the presented framework and its use as an effective Decision Support System in clinical practice in order to enhance patients' outcome. It is difficult, however, to go from a cross-sectional, retrospective study to a prospective interventional clinical trial. Temporal modelling is one of the missing major gaps of our current corpus of research. Risks of cardiovascular diseases are not a single number; risks change over the course of time as the patient ages and develops additional conditions and/or treatments. The current model only uses one clinical and imaging data to make a prediction; it does not include disease evolution. For instance, a borderliner with a normal echo at baseline might over a period of five years develop left ventricular hypertrophy and high cholesterol, which can add up considerably to the risk. This evolution will not be reflected by a model only including base data and could in fact lead to the patient being underreported.

Conclusion

In this paper, the attention-based multimodal deep fusion framework for cardiac risk stratification has been introduced which dynamically fuse both medical imaging and clinical variables. The main novelty is a learnable attention mechanism, which dynamically assigns a scalar weight for each modality, to dynamically highlight the most relevant signals for each patient. To that end, extensive experiments performed on a multimodal dataset of structured EHR data coupled with echocardiography images, have shown the ability of this adaptive fusion to substantially outperform single-modality baselines as well as simple concatenation strategies. Compared to the models using only clinical information or only imaging information, and the concatenation baseline, this proposed model achieves good accuracy of 96.2% and AUC value of 0.98, respectively.

References

- Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Baltrušaitis, T., Ahuja, C., et al. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Bozyel, S., Şimşek, E., Koçyiğit, D., Güler, A., et al. (2024). Artificial intelligence-based clinical decision support systems in cardiovascular diseases. *The Anatolian Journal of Cardiology*.
- Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., et al. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- D'Agostino, R., Ramachandran, S., & Pencina, M. (2008). The framingham heart study. General cardiovascular risk profile for use in primary care. *Circulation*, 2008.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*.
- Graves, A. (2012). Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.
- Kingma, D., & Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114.
- Lee, Y., Cha, J., Shim, I., Park, W., Kang, S., Lim, D., et al. (2023). Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction. *Npj Digital Medicine*.
- Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.
- Minhas, A., Talha, K., Abramov, D., et al. (2024). Racial and ethnic disparities in cardiovascular disease-analysis across major US national databases. *Journal of the National Medical Association*.
- Qiu, Y., Guo, H., Wang, S., Yang, S., Peng, X., et al. (2024). Deep learning-based multimodal fusion of the surface ECG and clinical features in prediction of atrial fibrillation recurrence following catheter ablation. *BMC Medical Informatics and Decision Making*.
- Qu, C., Zhang, X., Lu, Y., Wang, Y., & Su, C. (2026). MAF-net: Multimodal cross-attention-based fusion network for cardiovascular disease classification. *Plos One*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*.

- Rubin, D. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*.
- Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*.
- Selvaraju, R., Cogswell, M., Das, A., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*.
- Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*.
- Shen, D., Wu, G., & Suk, H. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*.
- Tsai, Y., Bai, S., Liang, P., Kolter, J., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, J., Li, J., Wang, R., & Zhou, X. (2024). VAE-driven multimodal fusion for early cardiac disease detection. *IEEE Access*.
- Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of*.