RESEARCH PAPER

HART-NET: A HYBRID ATTENTION, REINFORCEMENT LEARNING, AND TRANSFORMER-BASED MULTIMODAL ENSEMBLE FRAMEWORK FOR CARDIOVASCULAR DISEASE PREDICTION

[1*]Ajil D S Vins, [2]Sam Emmanuel W R,[3]Harriet Linda C

[1]Department of Computer Science, Nesamony Memorial Christian College, Marthandam, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, 627012, Tamil Nadu, India.*
Corresponding author email id: ajil_csa@nmcc.ac.in
[2]Department of Computer Science, Nesamony Memorial Christian College, Marthandam, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627012, Tamil Nadu, India.
Mail id:sam_emmanuel@nmcc.ac.in
[3]Department of Computing Technologies, SRM Institute of Science & Technology, Kattankulathur, Chengalpattu – 603203, Tamil Nadu, India.
Mail id:harrietc@srmist.edu.in

ABSTRACT

Global cardiovascular mortality is projected to increase from approximately 20.5 million deaths in 2025 to 35.6 million deaths in 2050. The existing Machine Learning (ML) and Deep Learning (DL) models are developed for individual data modalities and therefore may have limited capacity to integrate heterogeneous multimodal information, longitudinal temporal patterns, and clinically relevant contextual information. To address these challenges, this paper suggests HART-NET: a Hybrid Attention-based Reinforcement Transformer Network that integrates attention-based representation learning, reinforcement-learning-based feature weighting, Transformer encoding, graph neural networks, and ensemble learning for Cardiovascular Diseases (CVDs) classification. The five major phases that HART-

NET incorporates are (1) Data Preprocessing, where the UCI Heart Disease database, ECGs, clinical records, and wearable sensor data are cleaned, organized, and aligned to form a multi-modal data source, (2) Feature Engineering, where the proposed Convolutional-Attention Recurrent Deep Integration Architecture (CARDIA-Net) using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) is integrated with attention mechanisms and the proposed Evolutionary Blend-Crossover Genetic Optimizer (EvoBLX-GA) genetic optimization approach to extract and select relevant features; (3) Model Development, where Transformer-based attention mechanisms, reinforcement learning for dynamic feature weighting, and graph neural networks for modeling relationships are employed; (4) Ensemble Learning, which utilizes stacked ensemble methods and meta-learning techniques to obtain robust performance; and adaptive feature weighting, ensemble learning, and adaptability of models in the proposed HART-NET are reflected in its reported performance. HART-NET achieved accuracies of 98.55% and 98.91%, respectively, with corresponding G-Mean values of 98.62% and 98.96%. Five-fold cross-validation yielded a mean accuracy of 97.83% under both configurations. SHAP analysis identified influential clinical features contributing to model predictions. These findings demonstrate the potential of HART-NET for cardiovascular disease classification. Further external validation on diverse, independent clinical cohorts and quantitative explainability assessments are required to establish its generalizability and clinical applicability.

Keywords: *Cardiovascular Disease Prediction, Electrocardiography, Multimodal Data Integration, Transformer, Genetic Optimization, Graph Neural Network and Ensemble Learning.*

How to cite this article: 1*Ajil D S Vins 2Sam Emmanuel W R3Harriet Linda C. HART-NET: A HYBRID ATTENTION, REINFORCEMENT LEARNING, AND TRANSFORMER-BASED MULTIMODAL ENSEMBLE FRAMEWORK FOR CARDIOVASCULAR DISEASE PREDICTION. Int J Drug Deliv Technol. 2026;16(90s):1166-1183. DOI: 10.25258/ijddt.16.90s.115.

INTRODUCTION

Cardiovascular diseases (CVDs) have remained the leading cause of death worldwide. Recent WHO estimates indicate that CVDs account for approximately 19.8 million deaths annually based on 2022 mortality data. Early identification of individuals at elevated cardiovascular risk may support timely preventive and therapeutic interventions and may contribute to improved clinical outcomes [1]. The heterogeneity of cardiovascular manifestations, risk factors, comorbidities, and longitudinal clinical trajectories presents substantial challenges for conventional diagnostic and risk-assessment approaches [2]. The rapid growth of heterogeneous healthcare data, including Electronic Health Records (EHRs), wearable-device measurements, genomic information, physiological signals, and medical images, has created opportunities for applying Artificial Intelligence (AI) and data-analytics techniques to predictive healthcare [3]. In this study,

the proposed HART-NET integrates attention mechanisms, reinforcement learning, Transformer-based representation learning, graph neural networks, and multimodal ensemble learning to develop a unified framework for cardiovascular disease prediction [4-5].

The application of Machine Learning (ML) and Deep Learning (DL) to cardiovascular disease prediction has advanced substantially in recent years [6]. Traditional structured datasets, including the UCI Heart Disease dataset, have been extensively evaluated using Support Vector Machines (SVMs) and other conventional machine-learning methods, although reported performance varies substantially across datasets and evaluation protocols [7]. More advanced architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also been applied to cardiac data, particularly time-series signals and medical images [8]. Transformer architectures, originally introduced for natural-language processing, have subsequently been adapted to biomedical and clinical data because of their ability to model long-range dependencies [9]. Attention mechanisms can provide information about the relative contribution of input representations to model computations; however, attention-based attribution should not be assumed to constitute complete clinical interpretability. Reinforcement-learning approaches have also been investigated for sequential decision-making and adaptive clinical risk-management tasks [10]. Models that include several weak learners have shown great effects in generalization improvement and overfitting reduction. However, despite the advances made with these models, most of them are still unimodal and hardly explain most issues regarding their limited real-time applicability to heterogeneous populations [11].

Many published models are designed for one or a limited number of modalities and therefore do not directly address the joint integration of EHRs, physiological signals, wearable measurements, imaging data, and other heterogeneous clinical information [12]. The limited transparency of many deep-learning models remains an important barrier to clinical adoption, particularly when predictions are used to support high-stakes decisions. While reinforcement learning provides a means of decision-making in a sequence, applying this method to continually changing patient trajectories needs well-identified states, actions, rewards, transition models, and clinically relevant feedback [13]. Although Transformer models can effectively capture long-range dependencies, their computational and memory requirements may limit real-time deployment unless the architecture and inference strategy are appropriately optimized. Most current studies are also plagued with small or biased datasets, reducing the generalizability of the findings [14]. The need for an integrated, explainable, and adaptive framework, like HART-NET, that capitalizes hybrid attention mechanisms and reinforcement learning feedback in a multi-modal ensemble for improved accuracy, dynamism, and personalization of heart disease prediction in the big data analytics era is, thus, critical [15].

The main contributions are given below:

- Initially, the proposed CARDIA-Net, a hybrid feature-extraction architecture that combines convolutional layers, LSTM units, and attention mechanisms to learn complementary spatial/local and temporal representations from heterogeneous cardiovascular data.
- The proposed EvoBLX-GA for evolutionary feature selection, using BLX-α crossover and Gaussian mutation within a defined chromosome representation to optimize the trade-off between predictive performance and feature-subset size.
- Integration of representation learning using Transformer, Graph Neural Networks (GNN), reinforcement learning for feature weighting and stacked ensemble learning within HART-NET, followed by evaluation of this architecture relative to the defined baselines.

1. LITERATURE REVIEW

In 2023, Noroozi *et al.,* [16] analyzed the impact of feature selection methods on machine learning algorithms for heart disease prediction. The study evaluated different feature selection techniques and examined their effects on the predictive performance of machine learning models for identifying heart disease. Feature selection reduced irrelevant and redundant attributes and supported the development of more efficient predictive models. However, the effectiveness of the prediction depended on the selected feature selection method and the characteristics of the dataset. Differences in patient populations, clinical variables, data quality, and disease characteristics could also affect the generalization of the evaluated models.

In 2023, Dileep *et al.,* [17] used a cluster-based bi-directional LSTM for heart disease prediction. It applied K-Means clustering to remove duplicates from both UCI and real-time datasets, and used C-BiLSTM for prediction, followed by comparison with other prediction methods. The key limitation depended on the dataset quality and preprocessing to achieve maximum performance. It was shown that C-Bi-LSTM achieved a total accuracy; it was compared against six traditional methods in terms of accuracy, sensitivity, and F1 score.

In 2024, Kothinti [18] used CNNs, RNNs, and SVMs in a myriad of medical applications derived through data obtained from EHRs, wearables, and genomic sources. Pattern recognition

features had been developed in a large-scale medical dataset for use in the early prediction of diseases and personalized treatment planning. The limitations included algorithmic bias, data privacy issues, and ethical troubles. Findings accentuated the improvement in early diagnosis, cost reduction in health care, and improved outcomes through precision medicine by AI.

In 2025, Sathya & Magesh [19] examined a multimodal DL framework consisting of HRVGT, STC-T, SMM Fusion, EFS-CP, and CHRC-Net that stratifies cardiovascular risk by combining retinal and physiological signals. The process incorporates feature extraction, fusion, evolutionary selection, and contrastive risk prediction. The limitation revolved around specialized multimodal data sources and such computational architecture.

In 2024, Houssein *et al.,* [20] employed transfer learning from transformer-based language models such as BERT, RoBERTa, BioClinicalBERT, XLNet, and BioBERT in heart disease risk factor extraction from unstructured EHRs. It fine-tuned the models over the annotated i2b2 clinical NLP dataset for automated semantic extraction and risk prediction. The primary limitation was the need to resort to annotated clinical text datasets with domain-specific pretraining. An evaluation showed that RoBERTa achieved the highest micro F1-score of 94.27%. Still, this improvement was not statistically significant compared to other models, confirming that the transformer-based transfer learning approach is effective for the early detection of heart disease.

In 2026, Rao *et al.,* [21] developed a Transformer-based survival model for predicting all-cause mortality in patients with heart failure using data from multiple cohorts. The approach employed Transformer-based modelling to learn complex relationships within longitudinal clinical information and estimate patient survival outcomes. The multi-cohort design supported evaluation across diverse patient populations and provided a broader basis for assessing the predictive capability of the model. However, Transformer-based survival modelling involved considerable computational and architectural complexity and required sufficiently large longitudinal datasets. Differences in clinical practices, patient characteristics, data distributions, and missing information across cohorts could also affect model generalization and prediction reliability.

In 2023, Xu *et al.,* [22] introduced CVEformer, an attention-based neural network for predicting Major Adverse Cardiovascular Events (MACE) in patients under peritoneal dialysis. The process involved learning the associations among the heterogeneous and temporal EHR features through self-attention, identifying significant risk factors, and estimating MACE probability. The disadvantage was that high dimensionality, heterogeneity, and time series make EHR data very complex. The model outperformed the prediction performance of existing methods when validated using a large multi-centre patient dataset.

In 2025, Makarov & Lipkovich [23] developed a transformer-based model using BERT and its variants in the encoder-decoder configuration to predict future diseases based on past EHR data. The process followed a two-stage training scheme starting with unsupervised pre-training of the encoder and finishing with fine-tuning for sequence prediction. A drawback was a trade-off between precision and recall across model variants. Finally, results showed that the first model achieved 46.32% precision and 17.80% recall, while the second model achieved 24.98% precision and 46.17% recall, thus proving the usefulness of either model.

In 2025, Qiu *et al.,* [24] introduced CVDLLM, which is an interdisciplinary framework for cardiovascular disease diagnosis that employs large language model methods with CNNs, RNNs, HRV analysis, GAT, and GSAT with the aim of automatic diagnosis with 12-lead ECG signals. The approach involved extracting the time-series features, modelling the graph-based spatial-temporal interaction, and embedding the spatial information semantically using pretrained LLMs. The limitation was the difficult integration of multimodal and multiscale data sources.

In 2025, Talaat [25] used CNNs for estimating the cardiovascular risk in patients based on their health indicators. The methodology adopted involved embedding categorical data, extracting spatial features through the CNN layered structure, and applying analytically tuned hyperparameters of the network with dense-feature layers for regularization. However, their major limitation was associated with improved interpretability of the models and ethical safeguarding mechanisms. The model produced an R^2 value of 0.994, thus superseding conventional methods and demonstrating remarkable promise in decision-making support for clinicians.

In 2025, Adeyi *et al.,* [26] proposed a framework called CardioNet+, which included MSTCN with attention for ECG/PPG data and a hybrid architecture of ResNet-50 with Vision Transformer (ViT) for chest X-ray images. Also applied were multi-scale temporal convolutions for capturing temporal dependencies and attention mechanisms for emphasising the features. They proved their mettle in a multimodal classifier, bringing together local and global structural features. Their native strengths were high diagnostic accuracy and interpretability, which were validated through an ablation study on multi-head attention, ResNet-50 and SMOTE. Their lack of validation across heterogeneous datasets and non-inclusion of real-time wearable integration are major disadvantages.

1.1. Problem Statement

Transformers and DL have made great strides in modelling CVDs. However, every existing approach has limitations: dealing with high-dimensional, heterogeneous, and temporal EHR data, algorithmic bias, reliance on annotated datasets, low interpretability of models, and ethical concerns. Such challenges hold back concrete efficacy in real-world and clinical settings. Therefore, there is a need for a multimodal deep-learning framework that can integrate heterogeneous cardiovascular data while addressing predictive performance, computational efficiency, interpretability, and generalizability.

Table 1: Summary of Key Studies on Heart Disease Prediction Using AI Techniques

Author & Year	Aim	Methods Used	Advantages	Disadvantages
Norozi *et al.,* (2023) [16]	To investigate the impact of feature selection methods on machine learning-based heart disease prediction	Feature selection with ML	Identifies key features, reduces dimensionality, improves prediction	Performance depends on the selected feature-selection method and machine learning algorithm
Dileep *et al.,* (2023) [17]	Heart disease prediction using DL	K-Means + Cluster-based BiLSTM (C-BiLSTM)	High accuracy: 94.78% (UCI), 92.84% (real-time); outperformed six methods	Dependent on data quality and preprocessing
Kothinti (2024) [18]	Early disease prediction via EHRs, wearables, and genomics	CNNs, RNNs, SVMs	Personalized treatment, early diagnosis, cost reduction	Data privacy, ethical concerns, algorithmic bias
Sathya & Magesh (2025) [19]	Risk stratification combining retinal & physiological data	HRVGT, STC-T, SMM Fusion, EFS-CP, CHRC-Net	91% sensitivity, 89% specificity, 0.97 AUC	Requires multimodal specialized data and complex architecture
Houssein *et al.,* (2024) [20]	Risk factor extraction from unstructured clinical records	Transfer learning: BERT, RoBERTa, BioBERT, XLNet, BioClinicalBERT	High F1-score (94.27% RoBERTa); automated semantic extraction	Relies on annotated domain-specific datasets
Rao *et al.,* (2026) [21]	To predict all-cause mortality in patients with heart failure using multi-cohort clinical data	Transformer survival model	Captures complex patterns, improves risk prediction, validated across cohorts	Transformer-based survival models require substantial computational resources and large, high-quality clinical datasets
Xu *et al.,* (2023) [22]	Predicting MACE in peritoneal dialysis patients	CVEformer (self-attention-based neural network)	Outperformed existing models; multi-centre validation	Complex EHR data: high dimensionality and heterogeneity
Makarov & Lipkovic	Future disease prediction using	BERT encoder-decoder; 2-stage training	Demonstrated flexibility in diagno	Trade-off between precisio

h (2025) [23]	past EHR data		stic accuracy	n and recall
Qiu *et al.,* (2025) [24]	Multimodal ECG-based automatic heart disease diagnosis	CNNs, RNNs, HRV, GAT, GSAT, LLMs	State-of-the-art accuracy, recall, F1 score on ECG signals	Difficult integration of multimodal, multiscale data sources
Talaat (2025) [25]	Cardiovascular risk prediction using personal health indicators	CNN with embeddings, dense layers, hyperparameter tuning	High R^2 = 0.994; robust against traditional methods	Low interpretability; ethical implementation concerns
Adeyi *et al.,* (2025) [26]	To develop a multimodal diagnostic framework for accurate CVD prediction	MSTCN+Attention for ECG/PPG and hybrid ResNet-50 + Vision Transformer (ViT) for chest X-rays	High diagnostic accuracy and interpretability validated by ablation study (multi-head, SMOTE, ResNet)	Limited validation on diverse datasets; lacked real-time wearable integration

Table 1 gives an overview of recent literature (2023-2026) that uses ML and DL techniques to predict heart disease, assess cardiovascular risks, and predict patient outcomes. The reviewed studies address diverse objectives, including feature selection for improved prediction, early disease detection, risk stratification, mortality prediction, major adverse cardiovascular event (MACE) prediction, and automated diagnosis using multimodal clinical data. The methods range from conventional ML and feature-selection approaches to advanced DL architectures, including BiLSTM, CNNs, RNNs, graph-based models, Transformer-based models, BERT variants, and multimodal attention-based frameworks. Several studies demonstrate improved predictive performance through the integration of heterogeneous data sources such as EHRs, ECG and PPG signals, retinal images, chest X-rays, physiological indicators, and genomic or wearable data. These approaches offer advantages such as improved prediction accuracy, early diagnosis, personalized risk assessment, automated clinical information extraction, and enhanced representation of complex medical patterns. However, several challenges remain, including dependence on high-quality and adequately annotated datasets, data heterogeneity and high dimensionality, limited interpretability, privacy and ethical concerns, algorithmic bias, computational complexity, and limited validation across diverse real-world populations. Overall, the reviewed literature indicates a clear shift from conventional ML-based prediction toward advanced DL, Transformer-based, and multimodal approaches, highlighting their potential for more accurate, personalized, and early cardiovascular disease detection and risk assessment.

2. METHODOLOGY

The study process aims at a systematic derivation and testing of the proposed HART-NET framework, namely, a Hybrid Attention-based Reinforcement Transformer Network, to perform heart disease prediction effectively and sensibly based on multi-modal health-related data. This section starts with a description of the dataset, which consisted of ECG signals, clinical records, and wearable sensor data. The methodology subsequently explains the five most significant stages of the HART-NET framework which are (1) Data Preprocessing and Integration, in which the heterogeneous data sources are cleaned and normalized, as well as synchronized to the same time; (2) Feature Engineering and Selection, in which the proposed CARDIA-Net using CNNs, LSTMs, attention models as well as the proposed EvoBLX-GA using genetic optimization are utilized to extract and select important features; (3) Base Model Development, where Transformer-based attention, reinforcement learning to dynamically weight features, Graph Neural Network structure modelling take part; (4) prediction and (5) Ensemble and Meta-Learning, where stacked ensembles enhance robustness and generalization;, which performs final disease prediction and evaluates performance through cross-validation and statistical metrics. The overall flowchart for the proposed HART-NET Model is shown in Figure 1.

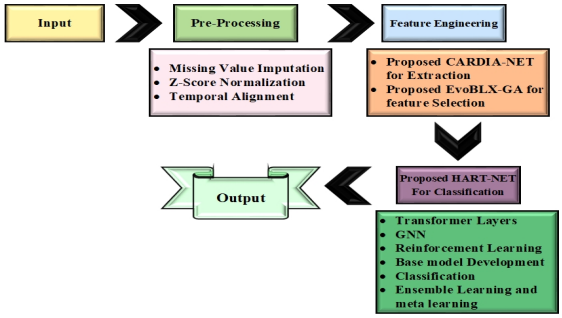
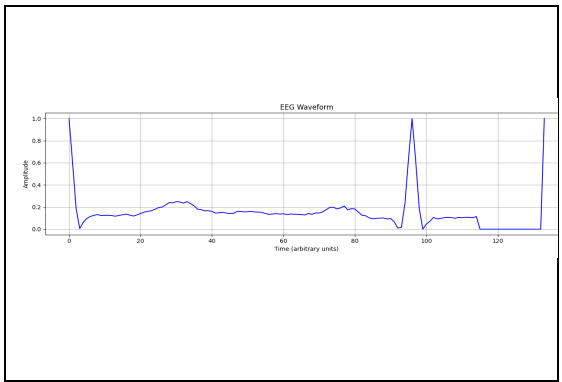


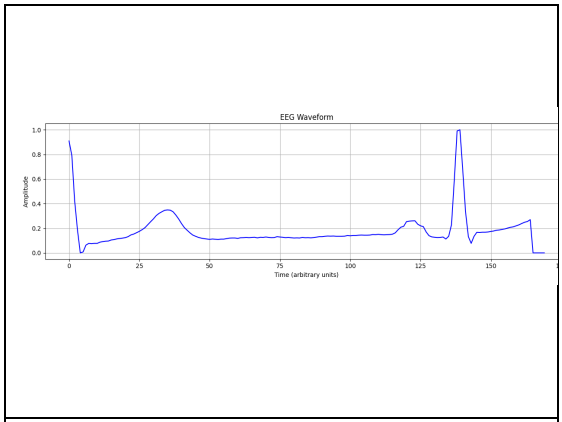
Figure 1: Overall flowchart for the proposed HART-NET Model

2.1. Dataset Description

To facilitate the strong classification of heart diseases and heartbeats, various datasets were obtained from the UCI Machine Learning Repository. Its 4 main tabulated datasets are based on the UCI Cleveland Heart Disease Database, Hungary, Switzerland, and the VA Long Beach, with 76 attributes, yet most of the works, including this one, use only 14 meaningful attributes. The target variable indicates whether heart disease is present on a scale of 0 no disease 1 to 4, and usually, it is simplified into classes 0 and 1: presence and absence of heart disease. In addition to this, two sets of ECG heart examination indicators were taken, which were obtained via the MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database. The data sets include pre-processed and segmented ECG data that can be used to train a deep neural network or to make use of transfer learning to categorise heartbeat type as normal or pathological. The synthetic wearable data refers to a collection of user attributes that deal with their health and activities obtained through smart devices. It contains demographics, sleep patterns, exercise, vital observations, lifestyle, environment, and indicators related to health. It also contains cross-sectional data that is time-stamped on user behaviour, health, and interactions with the environment.



a) Abnormal EEG waveforms



b) Normal EEG Waveforms

Figure 2 (a) and (b): Abnormal and Normal of EEG Waveforms

Figure 2 (a) and (b) illustrate the EEG waveforms, which are graphical representations of brain activity over time. Both plots show amplitude (arbitrary units) on the y-axis against time (arbitrary units) on the x-axis, with a blue line tracing the waveform. Both include a grid for easier reading of values. While having the same overall appearance and scale, the specific shapes and lengths of the peaks and troughs, especially the sharp, distinguishable spikes that resemble QRS complexes, differ between the two figures. This suggests that these recordings depict different brain electrical activity or possibly different segments of a longer recording.

2.2. Pre-Processing

Preprocessing is the first stage of the proposed pipeline and involves quality control, missing-value handling, normalization, temporal alignment, and modality-specific preprocessing to obtain a consistent multimodal representation. During this phase, ECG signals, clinical records, and wearable sensor data are systematically processed and aligned to form a unified multi-modal dataset suitable for subsequent modelling. This makes the dataset fit to train machine learning models with more accuracy and efficiency.

2.2.1. Missing Value Imputation

The process of replacing incomplete and missing data entries [27] with appropriate values to ensure the integrity of the dataset is called missing value imputation. This process helps maintain the consistency of the multi-modal inputs used to train the model. For numerical features with missing values, mean imputation is applied as per Eq. (1). To prevent information leakage, the imputation parameters are estimated exclusively from the training partition and subsequently applied unchanged to the validation and test partitions.

$$x_i = \begin{cases} x_i, & if\ x_i\ is\ present \\ \frac{1}{n}\sum_{j=1}^{n} x_j & if\ x_i\ is\ missing \end{cases} \quad (1)$$

2.2.2. Z-Score Normalization

Z-score normalization [28] is the technique employed to standardize features by the elimination of the mean and scaling to unit variance. This makes all the input features, which include ECG and sensor readings, comparable so that the model can learn effectively. For each continuous feature, the mean and standard deviation were calculated using only the training partition. The same parameters were then applied to the validation and test partitions. In this procedure, the formula is applied to transform each data point, where μ denotes the mean and σ indicates the standard deviation, as per Eq. (2).

$$z_i = \frac{x_i - \mu}{\sigma} \quad (2)$$

2.2.3. Temporal Alignment

Temporal alignment [29] refers to the alignment of multi-source data according to a shared timeline. Temporal alignment maps ECG, clinical, and wearable observations onto a common temporal reference. The manuscript specifies the reference time grid, sampling interval, interpolation or aggregation method, synchronization tolerance, and procedure used when one modality is unavailable at a given time point. This enables the model to learn cross-modal relationships correctly and generate better predictions, where $D(t)$indicates the aligned data vector at time t, and d denotes the number of combined features after integration, as per Eq. (3).

$$D(t) = [E(t), C(t), W(t)] \in R^{l \times d} \quad (3)$$

2.3. Feature Engineering

Feature Engineering is achieved by extracting valuable patterns of multimodal health data based on DL and optimization techniques. CNN layers are used to learn local representations from appropriately encoded ECG or imaging inputs, whereas LSTM layers model temporal dependencies in sequential physiological observations. Optimization based on genetics is used to select those features that help most in the optimization of the model and make it less complex.

2.3.1. Proposed CARDIA-NET for Extraction

The proposed CARDIA-NET framework is used to generate useful features of multimodal health data to predict heart disease correctly. It combines CNNs to extract spatial patterns from ECG signals and clinical representations. Temporal dependencies in time series data collected by sensors are modelled using LSTM networks. The attention mechanism is dynamic and points out the most important things in terms of time and modalities. This hybrid design guarantees high and attention-weighted feature extraction that is used in subsequent prediction.

Initially, let the input raw data $f^{(l)}$ is passed to the CNN [30], such as an ECG time series and a 2D image representation of a heartbeat. Then, the mathematical equations for the convolution operation are expressed as per Eq. (4), where $W^{(l)}$ denotes the Set of filter kernels in the layer l, $f^{(l-1)}$ indicates the input feature map from the previous layer, $*$ denotes the convolution operation, $b^{(l)}$ indicates the Bias term, and σ represents the activation function, typically ReLU. The filters slide across the input and compute dot products to detect local patterns such as the QRS complex in ECG, as per Eq. (4).

$$f^{(l)} = \sigma\left(w^{(l)} * f^{(l-1)} + b^{(l)}\right) \quad (4)$$

Moreover, the activation function of the convolution is passed through a non-linear function such as ReLU, as per Eq. (5). This introduces non-linearity and helps in learning complex patterns. Now, the pooling reduces the spatial dimensions while retaining important features. This makes the network invariant to small shifts and noise in the input. Then, after several convolution layers, the multi-dimensional output is flattened into a 1D vector. This is required for connecting to dense (fully connected) layers, as per Eq. (6). Additionally, the flattened vector is passed through one or more dense layers as per Eq. (7). This step performs high-level reasoning for prediction.

$$\sigma(x) = \max(0, x) \quad (5)$$

$$flattened\ Vector \in \mathbb{R}^n \quad (6)$$

$$y = \sigma(W_{dense}.x + b) \quad (7)$$

Additionally, LSTM [31] is yet another variant of RNN that learns to remember long-term relationships among sequences of data. An LSTM network contains a memory cell and three gating mechanisms such as the forget, input, and output gates that regulate the flow and retention of information across time steps. The same weight transfer also benefits the LSTM network in its analysis of time-related ECG signals or readings from wearable sensors to classify individuals with heart diseases. The function of neglect in this case determines what information is given up from the previous cell state c_{t-1}, where f_t discloses the results of the forget gate, x_t shows the information input at time t, h_{t-1} that of the previous hidden state, and σ indicates the sigmoid activation function as in the case of Eq. (8).

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (8)$$

In addition, the input gate decides new information to store as per Eq. (9), where i_t indicates the input gate output. Then, compute the candidate values, as per Eq. (10), where $\tilde{C}_t$ denotes the candidate's new cell content. Now, update the cell state to combine the forget gate and input gate as per Eq. (11), where $\odot$ denotes the element-wise multiplication, and c_t indicates the updated cell state. Finally, the output gate determines the output from the cell as per Eq. (12), and then computes the hidden state as per Eq. (13), where h_t indicates the output of the LSTM at time t.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (9)$$

$$\tilde{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (10)$$

$$c_t = f_t \odot C_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (12)$$

$$h_t = o_t \odot tanh(c_t) \quad (13)$$

Besides, the Attention mechanism [32] is a DL approach that allows models to capture the most salient parts of the input data when it comes to generating the prediction. It computes weighted representations of the input tokens or feature vectors according to their learned contextual relationships. The process enables the model to capture the contextual relationship and enhance task performance, such as classifying heart diseases using multilinear health data.

Here, the input feature vector $X \in \mathbb{R}^{n \times d}$ to generate the three separate projections, where n indicates the sequence length, and d denotes the feature dimension, $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ denotes the learnable weight matrices, and $Q, K, V \in \mathbb{R}^{n \times d_k}$ indicates the Query, Key, and Value matrices, as per Eq. (14).

$$Q = XW^Q, \; K = XW^K, \; V = XW^V \quad (14)$$

Then, the scaled dot-product attention scores are computed by taking the dot product of Q and the transpose of K, then scale by $\sqrt{d_k}$. This results in a similarity matrix of shape $n \times n$, showing how each position relates to others, as per Eq. (15). Now, normalize the scores using the softmax function to convert them into a probability distribution, as per Eq. (16), where α_{ij} indicates the attention position i gives to position j. Now, the weighted sum of values V is computed using the attention weight α, as per Eq. (17), where $n \times d_k$ indicates the output shape. This gives a context-aware representation of the input.

$$Scores = \frac{QK^T}{\sqrt{d_k}} \quad (15)$$

$$\alpha = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (16)$$

$$Attention(Q, K, V) = \alpha V = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (17)$$

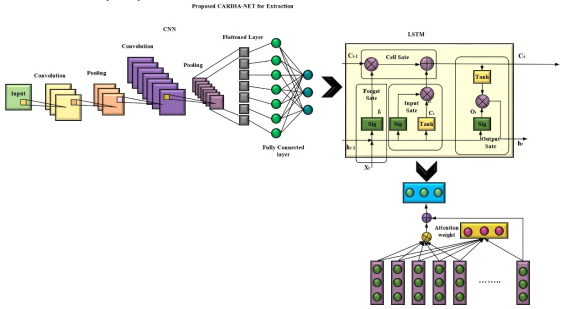


Figure 3: Proposed Architecture for CARDIA-NET for Extraction

Figure 3 depicts the proposed architecture called CARDIA-NET for feature extraction, which integrates CNNs, LSTMs, and attention mechanisms. In this method, spatial features from the input, such as ECG and clinical data, are first extracted by convolution, pooling, and fully connected layers within the CNN layers, then passed into an LSTM unit where temporal dependencies and sequential patterns are captured in its memory cells and gating mechanisms like forget, input, and output gates. Finally, an attention layer will dynamically assign weights to the outputs of the LSTM, emphasizing the most relevant features for downstream prediction, providing more precise and attention-weighted heart disease predictions.

2.3.2. Proposed EvoBLX-GA for Feature Selection

After the feature extraction, the dataset was transformed into a high-dimensional representation with a shape of (25,742, 415), capturing comprehensive information from multimodal inputs. The proposed EvoBLX-GA is a heterogeneous genetic algorithm meant to work efficiently and achieve optimality in selecting the features employed in heart disease prediction. It uses evolutionary techniques based on natural selection to refine subsets of features iteratively. BLX- a crossover is used to come up with new offspring, and Gaussian mutation is used to explore the combination of features. The fitness function is a trade-off between the maximization of prediction accuracy and reducing the number of features, maximizing performance, and interpretability. The roulette wheel selection method promises that elite members do have an increased chance of reproduction and, consequently, the convergence on the best feature sets that are rather quick.

Traditionally, the Genetic Algorithm (GA) [33] is a natural selection and genetics-based population-based optimization algorithm. It generates solutions in generations, and it recurs with operations such as selection, crossover, and mutation, with a fitness function used to rate the quality of any candidate. The goal is to maximize the fitness, which is frequently defined as per Eq. (18)

$$Fitness(x) = \alpha . Accuracy(x) - \beta . FeatureCount(x) \quad (18)$$

GA is a meta-heuristic technique that updates individuals through three stages: Selection, Crossover, and Mutation. Selection involves random pairs, Crossover creates new ones, Mutation replaces random pairs, and the population is updated until stop conditions are reached. The GA crossover operator is a basic modification of the single-point method, which involves randomly selecting two parents from the population and using a single point to divide information. This method creates new solutions by transferring values between the parents. However, for real code purposes, the blend crossover, also known as BLX-α, is more suitable,

extracting a portion from both parents, as per Eq. (19), where x_i^1 and x_i^2 indicate the elements taken from x^1 and x^2, and the α denotes the s positive value setting to 0.5.

$$x_i^1 = min(x_i^1, x_i^2) - \alpha d_i$$
$$x_i^2 = max(x_i^1, x_i^2) - \alpha d_i$$
$$d_i = |x^1 - x^2| \quad (19)$$

The mutation operator is responsible for investigating some specific solutions, akin to the crossover method. For example, according to Higashi and Iba, Gaussian mutation involves taking an element from the population and perturbing it with some random Gaussian number. The modified solution is then termed a mutated individual as calculated in Eq. (20), where x_{id} indicates the selected individual from the population, and σ denotes the random number generator that uses a Gaussian distribution with a standard deviation of σ = 0.1.

$$mutate(x_{id}) = x_{id} \times (1 + gaussian(\sigma)) \quad (20)$$

The selection operator extracts elements from a population for crossover and mutation manipulation. The roulette wheel is a common method, assigning a probability ps to each member, segmenting the population into regions, and computing the probability of selecting an element based on its fitness value, as per Eq. (21). Algorithm 1 presents the proposed EvoBLXGA for feature selection. Feature selection techniques were applied to reduce redundancy and enhance model efficiency, resulting in a refined dataset with 216 features while retaining critical information, thus reducing the input size to (25,742, 216) for optimal classification performance.

$$ps(a_i) = \frac{f(a_i)}{\sum_{i=0}^{n} f(a_i)}, j = 1,2, \dots, n \quad (21)$$

Algorithm 1: Proposed EvoBLX-GA for Optimal Feature Selection
Input: Dataset D with labelled samples and features, Population size N, Maximum generations G_{max}, Crossover rate p_c, mutation rate p_m, BLX parameter αGaussian mutation standard deviation $\sigma = 0.1$, Fitness function weight α_f, β_f **Output:** Best feature subset F_{best}
Steps:
Initial population Generate initial population $P = \{F_1, F_2, \dots, F_N\}$ with random feature subsets. Evaluate Fitness For each individual F_i, compute, as per Eq. (18). Repeat for G=1 to G_{max} Selection (Roulette Wheel) Select parent pairs (F_i, F_j) based on selection probability, as per Eq. (21). Crossover (BLX-α method) For each gene g in parents, as per Eq. (19). Mutation (Gaussian) For each gene g in a child, as per Eq. (20). Evaluate Offspring Fitness Repeat Step 2 for new offspring. Elitism and Replacement Retain top k elite individuals. Replace the rest with the best-performing offspring. Until the termination condition is met Return the best feature subset $F_{best} = arg \max_{f_i \in p} Fitness(F_i)$

2.4. Proposed HART-NET Framework for prediction

HART-NET is a comprehensive DL model designed to classify heart disease accurately and explainably. It combines transformers, GNN, reinforcement learning, and ensemble and meta-learning into a data-driven integrated framework. Initially, the input features from these modalities are first converted into fixed-length vectors for the transformer [34] using embedding layers, and to retain the order of the sequence, positional encodings are added to each token, forming an input matrix $X \in \mathbb{R}^{n \times d}$, where n denotes the number of time steps and d denotes the feature dimension.

Here, the input X is linearly projected into three matrices as queries $Q = XW^Q$, Key $K = XW^K$, and $V = XW^V$, where W^Q, W^K, W^V are learnable parameters, and d_k denotes the dimension of the projected space. These matrices allow the model to compute the importance of each token in the sequence relative to others using scaled dot-product attention, as per Eq. (22). The outcome of this operation produces an output that is weighted according to its attention for every token over all the other tokens in the entire sequence based on the learned relevance scores.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (22)$$

The results of the attention operation are then fed through a residual connection and normalization layer, ensuring stability and thus better training convergence, as per Eq. (23). Next, each position (token) is input into an independently processed feed-forward network (FFN) that incorporates a degree of non-linearity and a more profound transformation of it, as per Eq. (24), and finally, the FFN output is again passed through a second residual connection and layer normalization, as per Eq. (25). This final output $H \in \mathbb{R}^{n \times d}$ is stacked into deeper transformer layers or passed to downstream components like the Graph Neural Network or classifier.

$$\tilde{Z} = LayerNorm\big(X + Attention(Q, K, V)\big) \quad (23)$$

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (24)$$

$$H = LayerNorm\left(\tilde{Z} + FFN\left(\tilde{Z}\right)\right) \quad (25)$$

Moreover, GNN [35] uses a graph convolution to model their inter-feature dependencies, such as correlation between heart rate and blood pressure, where $\tilde{A} = A + I$ denotes the adjacency matrix with self-loops, $\tilde{D}$ indicates the diagonal degree matrix of $\tilde{A}$, $H^{(l)}$ represents the feature matrix at layer l,$W^{(l)}$ indicates the weight matrix, and σ represents the activation function, such as ReLU, as per Eq. (26).

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (26)$$

Then, reinforcement learning, such as dynamic feature weighting [36] using Q-learning, learns an agent to assign importance scores as per Eq. (27), where s_t indicates the state, a_t denotes the action, r_t indicates the reward, and γ denotes the discount factor. Additionally, the prediction uses the softmax as per Eq. (28), where z_i indicates the logit output from the last layer for the class i, and the mathematical formulation for the prediction per Eq. (29).

$$Q(s_t, a_t) = r_t + \gamma \max_a Q(s_{t+1}, a') \quad (27)$$

$$\hat{y}_i = softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (28)$$

$$Class = arg\max_i \hat{y}_i \quad (29)$$

Finally, the stacked ensemble learning and meta-learners are used to enhance robustness and generalization. Here, the multiple base learners are trained independently. Their prediction is expressed as per Eq. (30), where $f_j$indicates the j-th base prediction, and $\hat{y}_i^{(j)}$ denotes the prediction for the sample x_i. Then, the meta-learner takes base model outputs as input and learns to combine them as per Eq. (31), where $\mathcal{M}$ indicates that it is often a logistic regression and neural network that learns the optimal combination.

$$\hat{y}_i^{(j)} = f_j(x_i), \qquad j = 1,2,\dots,m \quad (30)$$

$$\hat{y}_i = \mathcal{M}\big(\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots\dots, \hat{y}_i^{(m)}\big) \quad (31)$$

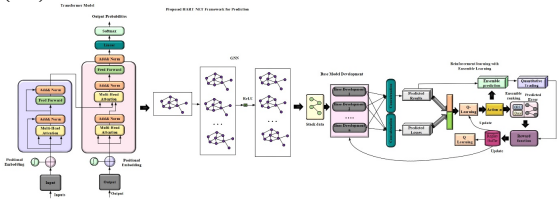


Figure 4: Proposed Architecture for HART-NET Framework for prediction

Figure 4 represents a blend of a Transformer model, GNN, RL, and ensemble learning in an intricate system for quantitative trading. Initial input data undergoes processing in a Transformer encoder, which creates embeddings that are transferred into the GNN. The GNN, consisting of layers of graph convolution using ReLU activation functions, appears to learn how to either make or transform data structured in the form of graphs. Output from the GNN, possibly combined with output from the Transformer, then enters the base Model development, where various base models are developed, classified results, and their losses are stacked. Stacked results go into an ensemble prediction module and also provide inputs into the prediction losses, which helps to train. Finally, a reinforcement learning loop based on Q-learning takes actions with ensemble prediction and updates its replay buffer from reward function responses with the end goal of optimizing quantitative trading through action check and ensemble ranking based on classified trends.

3. RESULT & DISCUSSION

The experimental results of the proposed HART-NET framework are presented and discussed in this section. The analysis in the work will be dedicated to the predictive performance, the effectiveness of the framework, and the computational efficiency of the proposed approach through the use of various performance metrics. The proposed framework will be tested against existing deep learning algorithms and machine learning algorithms to determine the effectiveness of the

HART-NET framework in predicting cardiovascular diseases. The contribution of the individual parts of the HART-NET framework will be examined by carrying out ablation studies, and different split scenarios will be used to test the consistency of the framework's performance. The analysis will be done using the following performance metrics: accuracy, precision, sensitivity, specificity, F-measure, MCC, FPR, FNR, G-Mean, NPV, and computational time.

3.1. Experimental Setup

The proposed HART-NET framework was implemented in Python using TensorFlow and PyTorch. The experiments were carried out on a high-performance computing system that included an Intel Core i9 CPU, 64 GB of RAM, and an NVIDIA RTX 3090 GPU to accommodate big data analytics for ECG/PPG signal data, clinical historical records, and imaging data. HARTANET performance was compared against various models including LSTM, RNN, CNN, and Vision Transformer (ViT), which could provide different modelling strategies for time series and image data. Performance was assessed in terms of a wide array of metrics: Accuracy, Precision, Sensitivity, Specificity, F-measure, False Negative Rate (FNR), False Positive Rate (FPR), G-Mean, Negative Predictive Value (NPV), and Computational time. Moreover, benchmark models like GRU, ResNet, VGGNet, and ANN were tested under the same metrics while adding the Matthews Correlation Coefficient (MCC) to ensure a complete and fair judgment on predictive capability and efficiency.

3.2. Performance Metrics

The above formulas represent some of the fundamental performance metrics for classification model evaluation for heart disease prediction. Accuracy indicates the value of the total number of correct predictions from all cases, whereas precision is the number of real positives among the predicted positives. Sensitivity measures the capability of the model to accurately classify the real positive cases, while specificity measures its capability to identify actual negative cases. F1-score generally unites precision and recall into one balanced score. The MCC provides a full measure, even for unbalanced data. NPV reflects the ratio of true negatives to predicted negatives. FPR and FNR show the proportion of errors concerning correct positive and negative predictions, respectively. G-mean strikes a balance between sensitivity and specificity to gauge model stability. Finally, computational time is just the difference between the time from reception to the input and the time from reception to the output, which means how efficiently the model is in processing this data. Table 2 outlines the Performance metrics and their formulas.

Table 2: Performance metrics and their formulas

Metrics	Formula
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ Here, True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN),
Precision	$Precision = \frac{TP}{TP + FP}$
Sensitivity	$Recall = \frac{TP}{TP + FN}$
Specificity	$Specificity = \frac{TN}{TN + FP}$
F_measure	$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
MCC	$MCC = \frac{(TP \times TN) - (FP \times}{\sqrt{(TP + FP)(TP + FN)(TN +}}$
NPV	$NPV = \frac{TN}{TN + FN}$
FPR	$FPR = \frac{FP}{FP + TN}$
FNR	$FNR = \frac{FN}{TP + FN}$
GMean	$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$
Computational Time	$Computational\ Time = End\ Time - Start\ Time$

3.3. Performance Analysis

Figures 5 (a) and (b) represent the training and validation accuracy of the model in 100 events. Generally, both plots show increasing accuracy lines for both data sets from which the model was formed, indicating a good learning characteristic of the model. In Figure 5(a), the validation accuracy exceeds the training accuracy during part of the later training period. This pattern does not, by itself, demonstrate underfitting; it may reflect stochastic variation, regularization effects, data-split characteristics, or differences between the training and validation data distributions. In the case of Figure 5 (b), there is a dashed green line for validation accuracy: training and validation accuracies follow each other much more closely over the epochs, with validation accuracies often being slightly lower than or very similar to training accuracy, serving as a fairly good indicator of good generalized without severe overfitting or underfitting.

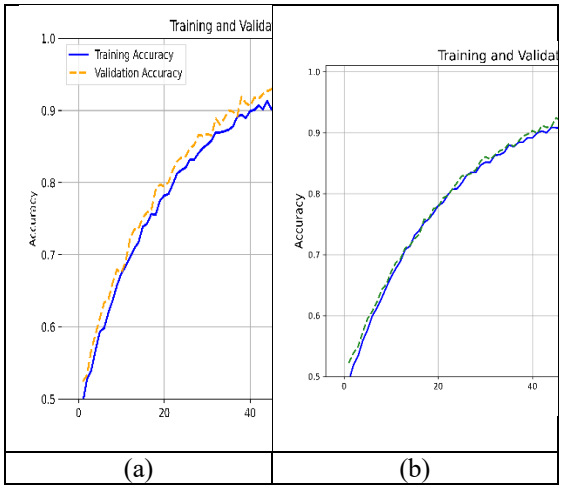


Figure 5: Training and Validation Accuracy Over Epochs for the Proposed Model (a) 70:30% and (b) 80:20%

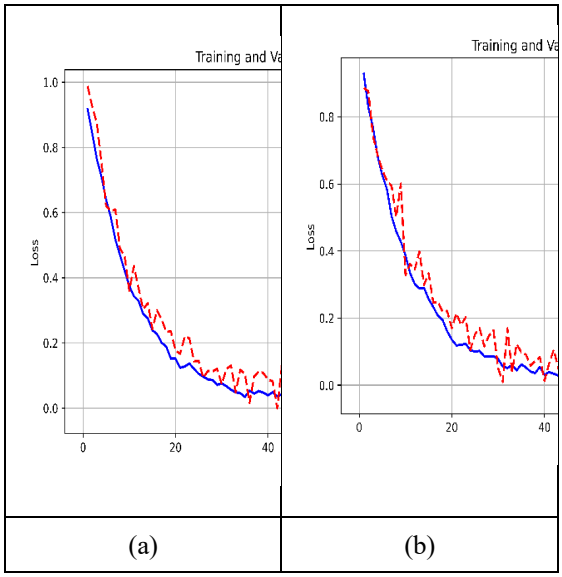


Figure 6: Training and Validation Loss Over Epochs for the Proposed Model, (a) 70:30% and (b) 80:20%

Figure 6 (a) and (b) depict the Training and Validation Loss Over Epochs for 70:30 and 80:20%, respectively, and display the trend of training and validation loss for a model over 100 epochs. In both graphs, the blue solid line represents the training loss, and the red dashed line represents the validation loss. Initially, both training and validation losses rapidly decrease, indicating that the model is effectively learning from the data. However, after approximately 20-30 epochs, the training loss continues to decrease and stabilizes at a very low value, close to zero, while the validation loss, although decreasing, starts to fluctuate and remains consistently higher than the training loss. This divergence between training and validation loss, where training loss is significantly lower than validation loss, suggests that the model might be overfitting to the training data, meaning it is performing well on the data it has seen but struggling to generalize to new, unseen data. The increasing separation between training and validation loss after approximately 20-30 epochs indicates a generalization gap and highlights the potential benefit of early stopping or regularization.

3.4. Comparative Metrics

From the two train-test split schemes 70:30% and 80:20%, the model proposed in this study is tested and compared with LSTM [17], RNN [20], CNN [25], and Vision Transformer (ViT) [26] based on different measures of classifier performance are summarised in Table 3. For the 70:30% scheme, the proposed model performs with accuracy of 98.55, precision of 98.37, sensitivity of 98.80, specificity of 99.60, F-measure of 98.58, MCC of 97.90, and NPV of 99.56 and for the 80:20% scheme, accuracy of 98.91, precision of 99.27, sensitivity of 98.85, specificity of 99.69, F-measure of 99.04, MCC of 98.44, and NPV of 99.68. The false positive rate (FPR) and false negative rate (FNR) are also lowest in the proposed model that measure the proportion of failing positive prediction and missing in the positive prediction respectively of an event in both schemes (FPR 0.40 and FNR 1.20 for the 70:30% scheme and FPR 0.31 and FNR 1.15 for the 80:20% scheme). The lowest values of FPR and FNR imply that the proposed model estimate fewer false-positive predictions and missing positive cases in comparison to the baseline models. In general, the proposed model shows an overall consistent and sustained prediction performance with high reliability and decreased errors in both train-test split configurations.

Table 3: Comparative Performance of the Proposed Model over Existing Models for 70:30% and 80:20%

Proposed Model Over Existing Models for 70:30%					
Metrics	**LSTM [17]**	**RNN [20]**	**CNN [25]**	**ViT [26]**	**Proposed**
Accuracy	93.84	91.67	93.48	94.2	98.55
Precision	92.87	90.95	92.64	93.79	98.37
Sensitivity	92.27	87.99	92.11	93.34	98.8
Specificity	98.41	97.8	98.3	98.47	99.6
F_measure	92.39	89.01	92.19	93.38	98.58
MCC	91.17	88.04	90.67	91.71	97.9
NPV	98.22	97.62	98.09	98.26	99.56
FPR	1.59	2.2	1.7	1.53	0.4
FNR	7.73	12.01	7.89	6.66	1.2
Proposed Model Over the Existing Model for 80:20%					
Accuracy	94.57	92.39	94.02	94.57	98.91

Precision	93.96	91.82	92.64	94.07	99.27
Sensitivity	94.55	90.92	93.59	95.08	98.85
Specificity	98.58	97.98	98.48	98.58	99.69
F_measure	94.2	91.13	93.03	94.48	99.04
MCC	92.2	89.12	91.43	92.21	98.44
NPV	98.43	97.77	98.31	98.41	99.68
FPR	1.42	2.02	1.52	1.42	0.31
FNR	5.45	9.08	6.41	4.92	1.15

3.5. Ablation Study Analysis

Table 4 shows the ablation study conducted on each aspect of feature engineering, the Transformer, reinforcement learning and the GNN and their respective effects on proposed model for the 70:30% and 80:20% train-test split configurations. 70:30% table, complete model yields the following results: 98.55 accuracy, 98.37 precision, 98.80 sensitivity, 99.60 specificity, 98.58 F-measure, 97.90 MCC and 99.56 NPV. Similarly, 80:20% table the proposed model achieves: 98.91 accuracy, 99.27 precision, 98.85 sensitivity, 99.69 specificity, 99.04 F-measure, 98.44 MCC and 99.68 NPV. Across all metrics, complete model performs best with 70:30% model having the best accuracy-related metrics and 80:20% model having the best sensitivity, specificity, F-measure, MCC and NPV values. For both train-test splits, the proposed model reports the lowest FPR, 0.40 and 0.31 respectively. However, false negative rate for the proposed model are 1.20 and 1.15 respectively, above those of the corresponding ablated variants and so results do not demonstrate a reduction of both types of errors. As a whole, the ablation results support the ability of integrated architecture to produce strong overall classification performance for both configurations while emphasizing the importance of examining FPR and FNR values separately.

Table 4: Ablation Study for the Proposed Model over 70:30% and 80:20%

Proposed Model Over 70:30%					
Metrics	**Without Feature Engin**	**Without Transformer**	**Without Reinforcement**	**Without GNN**	**Proposed**
Accuracy	97.23	97.65	97.89	98.12	98.55
Precision	96.87	97.12	97.35	97.8	98.37
Sen	97.35	97.54	97.76	98.	98.8
Spe	97.51	97.89	98.1	98.	99.6
F_m	97.1	97.41	97.68	97.	98.58
MCC	96.35	96.9	97.12	97.34	97.9
NPV	97.0	97.25	97.43	97.7	99.56
FPR	0.71	0.62	0.55	0.49	0.4
FNR	0.65	0.58	0.52	0.45	1.2
Proposed Model Over 80:20%					
Accuracy	97.34	97.86	98.23	98.54	98.91
Precision	96.82	97.27	97.65	98.12	99.27
Sensitiv	97.	97.68	98.0	98.4	98
Specifi	97.	97.89	98.1	98.61	99
F_meas	96.	97.43	97.8	98.3	99
MCC	96.24	96.75	97.36	97.81	98.44
NPV	97.13	97.52	97.9	98.36	99.68
FPR	0.63	0.57	0.49	0.38	0.31
FNR	0.59	0.51	0.41	0.32	1.15

3.6. Five-Fold Cross-Validation

The five-fold cross-validation results of the proposed model over 70:30% and 80:20% training and testing splits are shown in Table 5. It can observe that the proposed model with the 70:30% training:testing split yields a mean accuracy, precision, sensitivity, F-measure, MCC, FPR, and FNR of 97.83 0.57, 97.54 0.60, 95.67 0.54, 96.52 0.56, 96.85 0.83, 0.58 0.15 and 4.33 0.54, respectively, over 70:30% training:testing split. On the other hand, when 80:20% training:testing split is used, a mean accuracy, precision, sensitivity, F-measure, MCC, FPR, and FNR of 97.83 0.86, 97.55 1.08, 94.52 1.84, 95.76 1.62, 96.86 1.24, 0.59 0.22 and 5.48 1.84, respectively, are obtained. The corresponding mean NPV, FPR, and FNR of the proposed model are 99.41 0.17, 0.59 0.22, and 5.48 1.84 with 80:20% training:testing split. These indicate that both data splits give the same mean accuracy, i.e. 97.83. In addition, under the 70:30% training:testing split, most of the metrics achieved higher mean values with lower standard deviations than those obtained using 80:20% training:testing split. Moreover, it can be found that the specificity and NPV are higher than 99% over both training:testing splits. This confirms that the proposed method is capable of correctly classifying

positive and negative samples. Overall, the proposed model exhibits similar classification results over the five folds. However, a higher standard deviation of the accuracy, precision, sensitivity, F-measure, MCC, and error rate can be observed in results obtained using 80:20% training and testing split.

Table 5: Five-Fold Cross-Validation for the Proposed Model

70:30% Split Five-Fold Cross-Validation						
Metrics	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean ± SD**
Accuracy	98.55	97.46	98.19	97.1	97.83	97.83 ± 0.57
Precision	98.37	97.28	97.77	96.75	97.54	97.54 ± 0.60
Sensitivity	96.46	95.07	95.86	95.28	95.69	95.67 ± 0.54
Specificity	99.6	99.32	99.52	99.23	99.42	99.42 ±
F-measur	97.33	96.08	96.72	95.92	96.53	96.52 ±
MCC	97.9	96.32	97.37	95.81	96.85	96.85 ± 0.83
NPV	99.61	99.32	99.53	99.17	99.4	99.41 ± 0.17
FPR	0.4	0.68	0.48	0.77	0.58	0.58 ± 0.15
FNR	3.54	4.93	4.14	4.72	4.31	4.33 ± 0.54
80:20% Split Five-Fold Cross-Validation						
Accuracy	98.37	96.74	97.83	98.91	97.28	97.83 ±
Precision	98.36	96.32	97.56	98.85	96.65	97.55 ±
Sensitivity	95.51	91.52	95.09	96.29	94.18	94.52 ±
Specificity	99.57	99.13	99.41	99.68	99.28	99.41 ±
F-measur	96.76	93.23	96.16	97.41	95.26	95.76 ±
MCC	97.65	95.29	96.85	98.43	96.06	96.86 ±
NPV	99.57	99.11	99.42	99.74	99.29	99.43 ±
FPR	0.43	0.87	0.59	0.32	0.72	0.59 ±
FNR	4.49	8.48	4.91	3.71	5.82	5.48 ±

3.7. SHAP Analysis

Figure 7 shows individual feature SHAP values for the proposed model on the heart disease dataset. The features are ordered from top to bottom by their overall importance. Features at the top of the plot tend to have more positive or negative effect on the feature prediction as represented by the width of the distribution in the show. As the figure shows, cp, dataset, exang, and oldpeak tend to have large SHAP value distributions and thus tend to have a much larger overall effect on the proposed model. The age feature tends to have a large positive effect on model predictions, while thalach tends to have a positive effect on the proposed model. The features chol, sex, thal, and ca tend to have a less negative effect, while slope, trestbps, fbs, and restecg tend to have smaller positive and negative effect on many observations. The x-axis in the plot are the SHAP values representing how much a given feature pushes the propposed model toward the Target class. Blue coloration indicates a lower feature value and pink/red a higher feature value. For example, oldpeak and exang tend to have mostly positive (pushed the model toward the positive class) SHAP values with larger values being associated with larger pink shades, while the high values of age tend to have larger negative SHAP values being more associated with the blue shades. The plot shows several features where the effects depend on the value of the variable and are not merely linear relationships with the target. The proposed model can help identify those clinical and dataset features that have the most effect on predicted outcome.

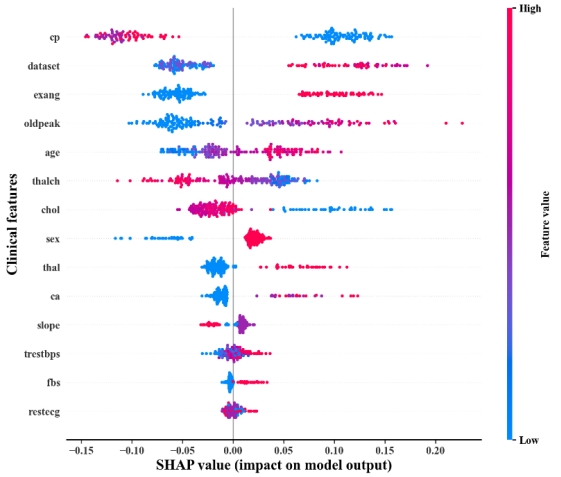


Figure 7: SHAP Summary Plot of Clinical Features and Their Impact on Model Predictions

3.8. Confusion Matrix

The confusion matrix of the proposed model with 70:30% training-testing split for 5 class (0-4). Figure 8(a) shows that the proposed model correctly predicts 121 class 0 instances, 79 class 1 instances,

33 class 2 instances, 31 class 3 instances, and 8 class 4 instances. Only 4 instances are predicted wrong: 1 class 0 instance is predicted as class 1, 1 class 0 instance is predicted as class 3, 1 class 1 instance is predicted as class 0, and 1 class 3 instance is predicted as class 2. Interestingly, all class 2 and class 4 instances are correctly classified, and out of class 0 and class 1 there is only 1 instance predicted wrongly each. In total, 272 correctly predicted instances out of 276 total instances for the proposed model with 70:30% training and testing splits. The accuracy rate is nearly 98.55%, which shows that the proposed model has high accuracy with relatively few misclassifications and low number of wrongly predicted instances. (b) Confusion matrix of the proposed model in 80:20% train-test split over five classes (labels 0-4). In the 80:20 % train-test split, 81, 53, 21, 21, 6 instances belonging to classes 0, 1, 2, 3, 4, respectively, are correctly identified by the proposed model while one instance in class 0 and 2, respectively, are identified as class 1 (off-diagonal entries are highlighted in the table). All instances belonging to classes 1, 3, and 4 are correctly identified by the proposed model. Only 2 instances, one from class 0 and one from class 2, are misclassified by the proposed model. The proposed model classifies all instances of classes 1, 3, and 4 correctly while one instance from classes 0 and 2 are misclassified, amounting to a total of 182 correctly classified instances of 184 with an accuracy of 98.91%. The accuracy of the proposed model in this case with a high concentration of outputs along the main diagonal of the table and with the off-diagonal entries small is compared to the 70:30 % case.

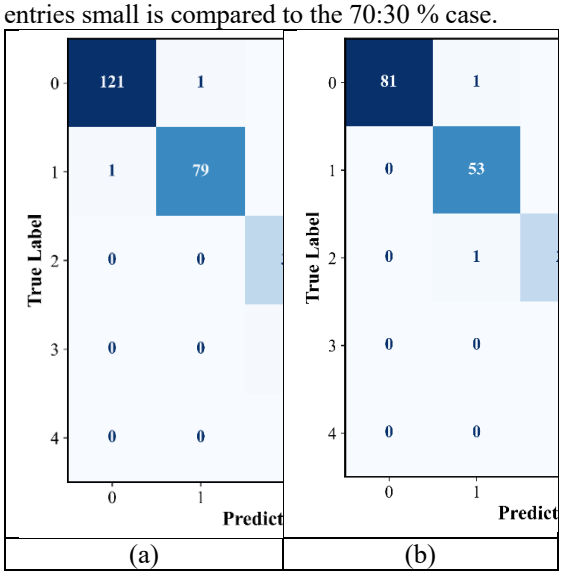


Figure 8: Confusion Matrix Analysis of (a) Training Data 70% and (b) Testing Data 80%

3.9. Analysis of Computational Time

Table 6 shows a comparison between the computational time of the LSTM, RNN, CNN, ViT and proposed models with train–test split of 70:30 and 80:20. The proposed model has the lowest computational time of 1.79 seconds for the 70:30 split compared to LSTM (2.03 s), RNN (2.28 s), CNN (2.16 s), and ViT (2.21 s). Likewise, the proposed model has a computational time of 1.84 seconds under the 80:20 split, which is better than LSTM (2.11 seconds), RNN (2.24 seconds), CNN (2.12 seconds), and ViT (2.25 seconds). The proposed model outperforms the fastest baseline (LSTM) by about 11.82% and 12.80% for the 70:30 and 80:20 splits respectively. The results show that the proposed model outperforms all the baseline models considered under both data-splitting configurations in terms of computational time, as it is required for the reported experiments.These results show that the proposed model has better computational efficiency than all the evaluated baseline models under both data-splitting configurations.

Table 6: Comparative analysis of computational time

Models	70:30%	80:20%
LSTM [17]	2.03	2.11
RNN [20]	2.28	2.24
CNN [25]	2.16	2.12
ViT [26	2.21	2.25
Proposed	1.79	1.84

4. CONCLUSION

This study presented HART-NET, a hybrid framework integrating attention mechanisms, reinforcement learning, Transformer-based learning, GNN, and ensemble strategies for multimodal cardiovascular disease prediction. The integration of feature engineering and complementary learning mechanisms enables the framework to capture complex relationships within cardiovascular data. Comparative evaluation demonstrated that HART-NET outperformed the LSTM, RNN, CNN, and Vision Transformer models under both 70:30 and 80:20 train-test splits. The model achieved accuracies of 98.55% and 98.91%, respectively, alongside high specificity, F-measure, MCC, and G-Mean values. Ablation experiments further highlighted the contribution of feature engineering, Transformer, reinforcement learning, and GNN components to overall classification performance, although their integration increased computational time. Five-fold cross-validation yielded a mean accuracy of 97.83% under both configurations, demonstrating generally consistent predictive performance. Additionally, SHAP analysis identified influential clinical features, improving model interpretability. Despite these promising results, further validation using larger, independent, and diverse clinical cohorts is necessary to establish generalizability and clinical applicability. Future work should investigate external validation, quantitative explainability, and privacy-preserving approaches to support reliable deployment in clinical environments.

DECLARATIONS:

FUNDING

On Behalf of all authors the corresponding author states that they did not receive any funds for this project.

CONFLICTS OF INTEREST

The authors declare that we have no conflict of interest.

COMPETING INTERESTS

The authors declare that we have no competing interest.

DATA AVAILABILITY STATEMENT

No new dataset was generated in this study.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCE

[1] Baghdadi, N. A., Farghaly Abdelaliem, S. M., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, *10*(1), 144.

[2] Kumar, A., Singh, K. U., & Kumar, M. (2023). A clinical data analysis based diagnostic systems for heart disease prediction using ensemble method. *Big Data Mining and Analytics*, *6*(4), 513-525.

[3] Mazumder, M. S. A. (2024). The Transformative Impact of Big Data in Healthcare: Improving Outcomes, Safety, and Efficiencies. *Global Mainstream Journal of Business, Economics, Development & Project Management*, *3*(03), 01-12.

[4] Adnan, M., Yi, Y., Wang, E., & Imtiaz, M. N. (2025). StackTrans–Multimodal Heart Disease Detection Using Stacked Transformer Fusion Framework. *IEEE Access*.

[5] Patra, R., Dutta, S., Roy, I. K., Basak, P., & Ghosh, A. (2025). Heart Disease Detection using Vision-Based Transformer Ensemble Models. *Procedia Computer Science*, *258*, 3554-3569.

[6] Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), 35-50.

[7] Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., & Algarni, A. (2024). Machine learning-based cardiovascular disease detection using optimal feature selection. *IEEE Access*, *12*, 16431-16446.

[8] Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., & Shaikh, A. (2023). A robust heart disease prediction system using hybrid deep neural networks. *IEEE Access*, *11*, 121574-121591.

[9] Jaltotage, B., Lu, J., & Dwivedi, G. (2024). Use of artificial intelligence including multimodal systems to improve the management of cardiovascular disease. Canadian Journal of Cardiology, 40(10), 1804-1812.

[10] Bizimana, P. C., Zhang, Z., Hounye, A. H., Asim, M., Hammad, M., & El-Latif, A. A. A. (2024). Automated heart disease prediction using improved explainable learning-based technique. *Neural Computing and Applications*, *36*(26), 16289-16318.

[11] Alam, M. A., Sohel, A., Uddin, M. M., & Siddiki, A. (2024). Big data and chronic disease management through patient monitoring and treatment with data analytics. Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems, 1(01), 77-94.

[12] Nelson, J., & Luther, D. (2025). Developing a Comprehensive Framework for AI-Driven Multimodal Data Integration in Heart Failure Management.

[13] Anderson, K. (2024). Explainable AI in Healthcare: Interpreting Predictions of Machine Learning Models for Heart Disease Diagnosis.

[14] Anderson, K. (2024). The Role of Data Preprocessing in Machine Learning Accuracy for Heart Disease Prediction Hybrid Models for Heart Disease Prediction: Combining Neural Networks with Traditional.

[15] Gayathri, R., Sangeetha, S. K. B., Mathivanan, S. K., Rajadurai, H., Malar, B. A., Mallik, S., & Qin, H. (2024). Enhancing heart disease prediction with reinforcement learning and data augmentation. *Systems and Soft Computing*, *6*, 200129.

[16] Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Scientific reports, 13(1), 22588.

[17] Dileep, P., Rao, K. N., Bodapati, P., Gokuruboyina, S., Peddi, R., Grover, A., & Sheetal, A. (2023). An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. *Neural Computing and Applications*, *35*(10), 7253-7266.

[18] Kothinti, R. R. (2024). Artificial Intelligence in Disease Prediction:

Transforming Early Diagnosis and Preventive Healthcare.

[19] Sathya, K., & Magesh, G. (2025). Multimodal Deep Learning for Cardiovascular Risk Stratification: Integrating Retinal Biomarkers and Cardiovascular Signals for Enhanced Heart Attack Prediction. *IEEE Access*.

[20] Houssein, E. H., Mohamed, R. E., Hu, G., & Ali, A. A. (2024). Adapting transformer-based language models for heart disease detection and risk factors extraction. *Journal of Big Data*, *11*(1), 47.

[21] Rao, S., Ahmed, N., Salimi-Khorshidi, G., Yau, C., Su, H., Conrad, N., ... & Rahimi, K. (2026). A transformer-based survival model for prediction of all-cause mortality in patients with heart failure: a multi-cohort study. npj Digital Medicine, 9(1), 118.

[22] Xu, Z., Xu, X., Zhu, X., Niu, K., Dong, J., & He, Z. (2023). Attention-Based Deep Learning Model for Prediction of Major Adverse Cardiovascular Events in Peritoneal Dialysis Patients. *IEEE Journal of Biomedical and Health Informatics*, *28*(2), 1101-1109.

[23] Makarov, N., & Lipkovich, M. (2025). A transformer-based model for next disease prediction using electronic health records. *The European Physical Journal Special Topics*, 1-10.

[24] Qiu, X., Wang, H., Tan, X., & Jin, Y. (2025). CVDLLM: Automated Cardiovascular Disease Diagnosis with Large-Language-Model-Assisted Graph Attentive Feature Interaction. *IEEE Transactions on Artificial Intelligence*.

[25] Talaat, F. M. (2025). Revolutionizing cardiovascular health: integrating deep learning techniques for predictive analysis of personal key indicators in heart disease. *Neural Computing and Applications*, *37*(1), 1-24.

[26] Adeyi, V., Xiaoling, Z., Uko, K., Kanjarawy, I., & Emmanuel, G. (2025). CardioNet+: Revolutionizing Heart Failure Diagnosis with Multi-modal Learning. *Available at SSRN 5132486*.

[27] Mbamba, C. K., Keymer, P., Alvi, M., Topalian, S. O., Din, F. U., & Batstone, D. J. (2025). Enhancing data quality in wastewater processes: Missing data imputation with deep Variational Autoencoders and genetic algorithms. *Computers & Chemical Engineering*, *199*, 109123.

[28] Putrawansyah, F., Rahayu, C., & Dhiniati, F. (2024). Application of particle swarm optimization to improve the performance of the k-nearest neighbor in stunting classification in South Sumatra, Indonesia. International Journal Of Education And Management Engineering, 6, 32-43.

[29] Chen, X., Ye, X., Xiong, X., Zhang, Y., & Li, Y. (2024). Improving the accuracy of wind speed spatial interpolation: A pre-processing algorithm for wind speed dynamic time warping interpolation. *Energy*, *295*, 130876.

[30] Liu, Y., Li, D., Zhao, J., & Liang, Y. (2024). Enhancing heart failure diagnosis through multi-modal data integration and deep learning. Multimedia Tools and Applications, 83(18), 55259-55281.

[31] Zhu, J., Liu, H., Liu, X., Chen, C., & Shu, M. (2025). Cardiovascular disease detection based on deep learning and multi-modal data fusion. Biomedical signal processing and control, 99, 106882.

[32] Yoon, T., & Kang, D. (2023). Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases. *Journal of Personalized Medicine*, *13*(2), 373.

[33] Fatima, S., Purushothaman, R., Jeyalakshmi, M. S., Gupta, R., & Sivakumar, N. (2024). Heart Disease Prediction based on Convolutional Neural Network Feature Genetic Algorithm Solutions. *Frontiers in Health Informatics*, *13*(7).

[34] Adnan, M., Yi, Y., Wang, E., & Imtiaz, M. N. (2025). StackTrans–Multimodal Heart Disease Detection Using Stacked Transformer Fusion Framework. *IEEE Access*.

[35] Li, J., Fan, L., Wang, X., Sun, T., & Zhou, M. (2024). Product demand prediction with spatial graph neural networks. *Applied Sciences*, *14*(16), 6989.

[36] Chen, Y., Gao, J., & Wu, J. (2024). Dynamic feature selection in medical predictive monitoring by reinforcement learning. *arXiv preprint arXiv:2405.19729*.

[37] Khan, H., Javaid, N., Bashir, T., Akbar, M., Alrajeh, N., & Aslam, S. (2024). Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BlCVDD-Net. IEEE Access, 12, 109230-109254.

[38] Ahmed, M. R., Aziz, A., Manik, M. M. H., & Habib, M. A. (2025). Blockchain-aided comparative study of heart disease detection using machine learning-based approaches with an expanded dataset. Computers in Biology and Medicine, 195, 110611.

[39] Goswami, A. D., Bhavekar, G. S., & Chafle, P. V. (2023). Electrocardiogram signal classification using VGGNet: a

neural network based classification model. *International Journal of Information Technology*, *15*(1), 119-128.

[40] Mondal, S., Maity, R., & Nag, A. (2025). An efficient artificial neural network-based optimization techniques for the early prediction of coronary heart disease: comprehensive analysis. *Scientific Reports*, *15*(1), 4827.