A Two-Stage Hybrid Deep Learning Model for Heart Disease Prediction Using ECG and Echo Images and AO Based Feature Optimization

Bommaiah Boya[1], Dr.P.Devaraju[2]

1. Research Scholar in Computer Science & Technology, Sri Krishnadevaraya University, Ananthapuramu, Andhra Pradesh-515003, India.
2. Assistant Professor in Computer Science & Technology, Sri Krishnadevaraya University, Ananthapuramu, Andhra Pradesh-515003, India.

Correspondence: Bommaiah Boya

ABSTRACT

Detecting heart disease at an early stage is essential for lowering death rates and ensuring timely medical care. This study introduces a two-stage hybrid deep learning model that integrates Electrocardiogram (ECG) and Echocardiogram (Echo) images for accurate heart disease prediction. For both data types, features are extracted using advanced vision-based models such as Vision Transformer (ViT) and Swin Transformer. A metaheuristic approach Aquila Optimizer (AO) is employed for optimal feature selection. These refined features are classified using a fine-tuned Multi-Layer Perceptron (MLP). A comparative study showed that the Swin Transformer-based fusion outperformed ViT-based models, reaching 96.76% accuracy, 97.98% F1-score, 96.19% precision, and 99.84% recall. This approach demonstrates the effectiveness of combining multimodal imaging data with AO optimization for enhancing automated heart disease diagnosis systems.

Keywords: Cardiovascular diseases (CVDs), Electrocardiogram (ECG), Echocardiogram (Echo), Vision Transformer (ViT), Swin Transformer, Aquila Optimizer (AO), Multi-Layer Perceptron (MLP), heart diseases.

1. Introduction

According to the World Health Organization (WHO), cardiovascular diseases continue to be the leading cause of death worldwide. In 2019, WHO reported approximately 17.9 million deaths globally from cardiovascular-related conditions[2][3]. By 2022, cardiovascular conditions represented nearly 32% of all deaths worldwide, about 19.8 million cases, with around 85% linked to heart attacks and strokes. [1]. These diseases encompass a wide range of conditions affecting the heart and blood vessels, such as coronary artery disease, irregular heart rhythms, and heart failure. The prevalence of CVDs is increasing rapidly due to lifestyle choices, aging populations, and delayed detection, especially in low and middle-income nations [1][15]. Early and accurate detection of heart disease is critical to reducing fatal outcomes and facilitating prompt medical intervention. As a result, there is an increasing need for advanced, automated, and intelligent diagnostic systems that can assist clinicians in accurately identifying CVDs with minimal human error and high reliability [2]. CVDs are commonly diagnosed using techniques like electrocardiography (ECG), echocardiography, computed tomography (CT), angiography and magnetic resonance imaging (MRI). Classic diagnostic methods like electrocardiography [32] and echocardiography [9] are crucial in evaluating the heart's condition. ECG records the electrical activity of the heart and is effective in detecting arrhythmias and myocardial infarctions [11], while Echo uses ultrasound to visualize heart structures, offering detailed insights into heart chamber size, valve functions, and pumping capacity. Echocardiograms play a vital role in diagnosing and treating various heart conditions, including cardiomyopathy, heart failure, and valvular heart diseases like aortic stenosis and mitral regurgitation[2][3]. By

evaluating important factors such as ejection fraction and diastolic function, and employing Doppler techniques to assess blood flow and cardiac hemodynamics, echocardiograms aid in detecting irregularities in pressure and flow. Nevertheless, the interpretation of these imaging modalities necessitates specialized knowledge, which is typically scarce in settings with limited resources. Thanks to recent progress in artificial intelligence (AI) and deep learning, automated diagnostic models [10] have become increasingly popular in medical image analysis, leading to more precise and efficient disease prediction. Despite the existence of multiple diagnostic techniques [12], accurately and promptly identifying heart disease continues to pose a significant challenge. Traditional approaches heavily depend on expert interpretation, which can result in inconsistencies in diagnosis and place a greater burden on medical professionals. Additionally, while researchers have investigated deep learning techniques for medical imaging analysis, most studies have focused on a single modality, such as ECG or echocardiogram images, which restricts the predictive capabilities of the models. It is crucial to develop a comprehensive approach that incorporates elements from both imaging modalities to enhance the accuracy of predictions and facilitate better clinical decision-making. In this study, we focus on creating a multimodal deep learning framework that combines information from both ECG and Echo images. The process of feature extraction is performed using ViT [4] and Swin Transformer models [5] separately, and then classification is achieved using a fine-tuned multi-layer perceptron (MLP).The study also employs a metaheuristic optimization technique Aquila Optimizer (AO) to enhance the selection of the most relevant features before classification. This hybrid approach leverages AO's strong exploration and exploitation capability to navigate the feature space effectively. While single-modality models offer useful insights, they often lack the depth provided by multimodal data fusion. Experimental results reveal that models integrating multiple data modalities, coupled with optimized feature selection via AO significantly outperform those based on a single modality.

2. Literature Review

Progress in AI and deep learning has significantly improved medical image analysis, allowing quicker and more precise diagnoses. This review highlights the growing role of deep learning in heart disease prediction, particularly through ECG and Echocardiogram imaging. While many studies apply CNNs, RNNs, ResNet, VGG, and EfficientNet family on individual modalities to extract features from ECG, Echo images and train the model to predict heart diseases, the use of multimodal fusion remains limited. Models like ViT and Swin Transformer demonstrate strong potential but are frequently limited to single-modality use. Moreover, existing machine learning and deep learning fusion approaches rarely incorporate advanced optimization or automated feature selection, indicating a gap in fully integrated and optimized multimodal frameworks.

Muhammad Raoof et al., [2] used the latest VGG16 deep learning model to automatically detect heart disease from echocardiogram images, and they achieved an accuracy of 94.92%. Muhammad Mahtab et al., [3] developed a deep learning model called EfficientNetB3, which was made to identify critical congenital heart disease (CCHD) from echocardiogram pictures. This model got an accuracy of 94.53%.

Chunyu Fan et al., [4] introduced a new model named ViT-FRD, which mixes features from a Vision Transformer (ViT) and a CNN using a process called knowledge refinement. This model showed very high precision, with a success rate of 96.90%.

Zeynep Hilal Kilimci et al., [5] did a study to detect heart disease using ECG images. They used vision transformer models like Google's ViT, Microsoft's Beit, and Swin-Tiny. The Beit model performed the best, achieving an

accuracy of 95.9%, which was better than the other two models.

Jingyuan Yi et al., [25] made an improved transformer model aimed at boosting heart disease prediction accuracy. They used particle swarm optimization (PSO) to reach a classification accuracy of 96.5%.

Tb Ai Munandar [21] proposed an MLP model that used different activation functions like tanh, logistic, and ReLu. The MLP model with the tanh activation function performed better than the logistic and ReLu models in terms of accuracy. When trained using tanh and k-fold cross-validation, the MLP model achieved a classification accuracy of 78.8%.

Alexey Dosovitskiy et al., [6] did a study on using transformers for image recognition. They treated an image as a set of patches and used a transformer encoder, which is typically used for natural language processing, to analyze the image. The Vision Transformer (ViT) performed much better than leading CNNs and needed fewer computational resources to train.

Mahmoud Khalil et al., [7] did a detailed review of past research on vision transformers for image classification, organizing the models in the order they were developed.

Koki Nakanishi et al., [9] conducted a review on the possible heart-related causes of stroke and evaluated how important echocardiography is in clinical settings.

Priya Dubey et al., [10] wrote a paper that gives a detailed overview of various deep learning methods used to predict heart disease, including CNNs, RNNs, and hybrid models. They also discussed current diagnostic and predictive techniques used in clinical evaluations and imaging methods.

Lerina Aversano et al., [11] used ECG images linked to different heart issues to predict heart disease with deep learning tools. They found that the CNN-2D model was able to correctly identify heart disease in ECG images about 91% of the time.

Numerous studies have been widely applied CNNs to analyze ECG data for arrhythmia classification and to Echocardiograms for segmenting heart chambers and detecting cardiomyopathies. However, their inability to model long-range dependencies limits their effectiveness in complex medical imaging. To address this, our study employs transformer-based architectures like Vision Transformer (ViT) and Swin Transformer—which have shown promising results in recent biomedical imaging research by capturing global context and deeper feature representations.

3. Proposed Methodology

The literature review emphasizes that most existing studies rely on single-modality ECG or Echo imaging data and often lack sophisticated feature fusion or optimization methods. To address these gaps, our research introduces a novel dual-modality framework that combines ECG and Echo features extracted via Vision Transformer (ViT) and Swin Transformer. A metaheuristic algorithm, Aquila Optimizer (AO) is used for selecting the most relevant features prior to classification. The proposed pipeline includes imaging datasets acquisition, pre-processing of datasets, features extraction from datasets, multimodal features fusion, AO based feature selection, and classification using a fine-tuned MLP classifier. The proposed approach enhances diagnostic performance by leveraging both structural and electrical heart characteristics for more accurate heart disease prediction.

3.1 Datasets Acquisition

In this study, we utilize two real-time imaging datasets as Electrocardiogram (ECG) and Echocardiogram (Echo) and these are collected manually from Government General Hospital [35]. Each dataset comprises 4500 images, including 2500 normal and 2000 abnormal cases, representing the same individuals across both imaging modalities. Table 1 presents the distribution of normal and abnormal images of each dataset, and also figure 1, figure 2 illustrates sample normal and abnormal ECG and Echo images from both categories.

Table 1: Presents total images in ECG, Echo datasets

	ECG images	Echo images
Normal images	2500	2500
Abnormal images	2000	2000
Total images	4500	4500

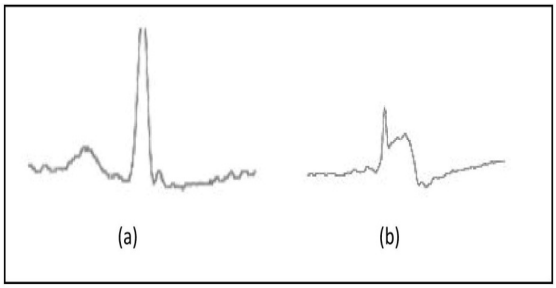

Figure 1: Represents (a) normal ECG image (b) abnormal ECG image

Figure 2: Represents (a) normal Echo image (b) abnormal Echo image

3.2 Datasets Pre-processing

In this study, we utilize Vision Transformer (ViT) and Swin Transformer to extracting features from ECG and Echo images, these two are deep learning models based on the Transformer architecture [6]. Since the original datasets contain images in varying sizes of 112×112, 192×192, and 224×224, so all images in the datasets are resized to 224×224 to meet models input requirements and converted to RGB colour format. To improve image quality, Gaussian filtering is applied for noise reduction, followed by z-score normalization using ImageNet mean and standard deviation to transform pixel values into the range of [0, 1]. Image augmentation techniques such as rotation, flipping, and brightness adjustment are employed to enhance data variability and model generalization. Heatmap is generated from the datasets confirm balanced class distributions, eliminating the need for further rebalancing. From the figure 3, it is evident that the datasets are already balanced, so there is no need to balance them again.

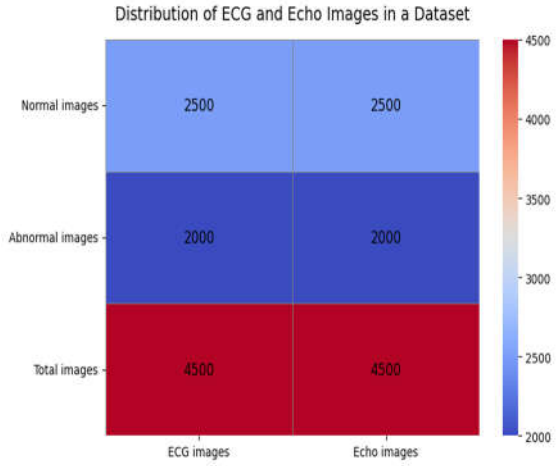

Figure 3: Represents total normal and abnormal cases in ECG, Echo datasets

3.3 Features Extraction, Fusion and Selection

To effectively capturing spatial and contextual patterns in ECG and Echo images, in this study, we employ Vision Transformer (ViT) [19][38] and Swin Transformer models for features extraction. ViT segments input images into 16×16 patches and applies global self-attention mechanisms[31] to extract deep features, utilizing its encoder blocks while excluding the built-in classifier as shown in figure4[6][7]. Swin Transformer adopts a hierarchical architecture with shifted window-based attention, enabling efficient local and global feature learning through progressive stages and patch merging as depicted in figure5[7]. In both models, only the feature extraction layers are used, while a separate fine-tuned MLP is employed for final heart disease classification.

Algorithm 1: Feature Extraction, Fusion, and Selection using AO metaheuristic

Input: ECG and Echo image datasets.

Output: Optimal fused and selected features

Steps:

Step 1: Image Data Acquisition

Load ECG and Echocardiogram image datasets

Step 2: Image Pre-processing

Resize all images to 224×224 pixels, convert grayscale images to RGB, normalize pixel values to range [0, 1], denoise using Gaussian filtering, perform data augmentation, convert images to tensors.

Step 3: Feature Extraction using ViT and Swin Transformer

ViT (Vision Transformer):

Used pre-trained ViT Base/16 and Input: 224×224, Patch size: 16×16

Extract global features from final encoder block.

Swin Transformer:

Used pre-trained Swin-Tiny model and Patch size: 4×4, embedding dim: 96

Extract hierarchical local + global features from final stage.

Step 4: Feature Fusion

Concatenate features:

ViT_ECG + ViT_Echo → ViT_Fused and Swin_ECG + Swin_Echo → Swin_Fused

Step 5: Feature Selection using Aquila Optimizer (AO)

Initialize population of binary vectors (agents), size N = 30.

Apply Aquila Optimizer (AO):

Perform exploration:

Expanded & narrowed exploration inspired by eagle hunting behavior.

Perform exploitation:

Narrowed & expanded exploitation to refine local optima.

Use fitness function:

Evaluate each agent (feature subset) using a classification score (e.g., accuracy using a temporary model) and Select optimal feature subset with highest fitness score.

Algorithm 1: Presents features extraction, fusion and features selection

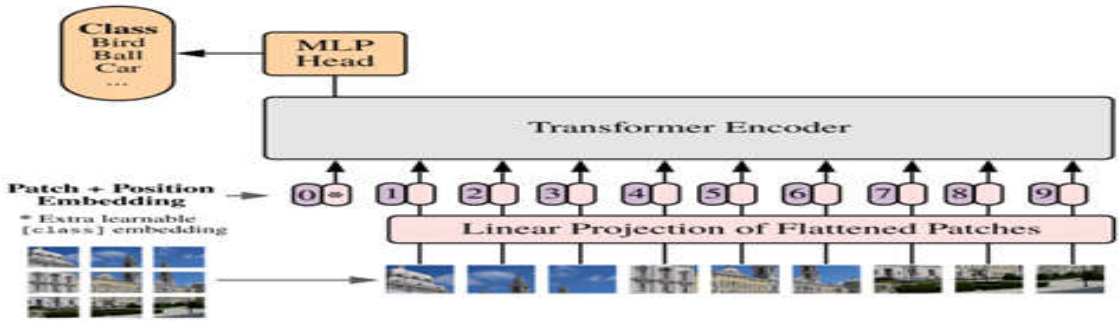

Figure 4: Represents ViT architecture

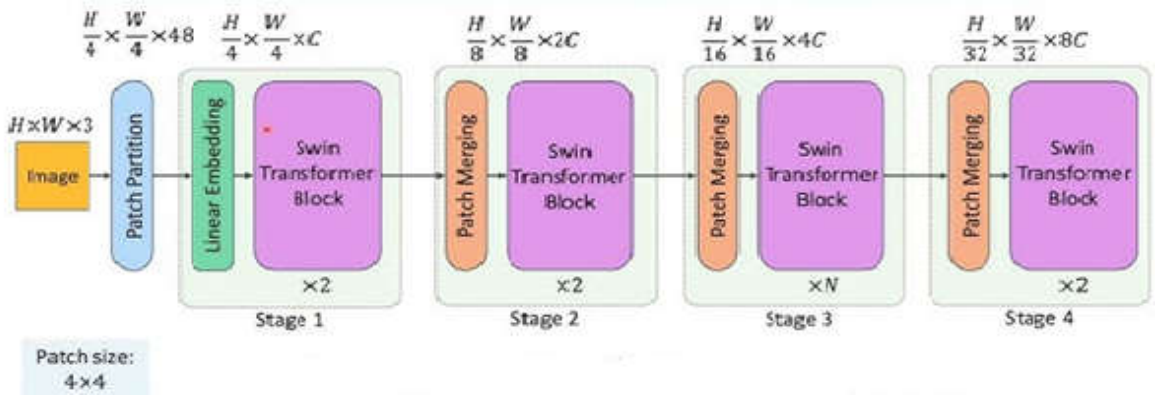

Figure 5: Represents Swin Transformer architecture

Algorithm 2: Model Building, Training, and Evaluation Using MLP

Input: Selected fused features

Output: Predicted class labels and performance metrics

Steps:

Step 1: Dataset Preparation

Split selected features and corresponding labels into train set (80%) and test set (20%).

Step 2: MLP Classifier Design

Input Layer: size = number of selected features
Hidden Layers: 128 → 64 → 32 neurons
Output Layer: 1 neuron with sigmoid activation
Activation: ReLU and Optimizer: AdamW
Dropout: 0.3 after each hidden layer
Loss: Binary Cross-entropy
Learning Rate: 0.00002 and Batch Size: 32
Epochs: 20, with Early Stopping (patience = 5)

Step 3: Model Training

Train MLP on the training data using selected features.

Step 4: Model Evaluation

Predict outcomes on the test set.

Compute performance metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC and Generate: Confusion Matrix, ROC Curve, and Precision-Recall Curve.

Algorithm 2: Presents MLP classifier design, training and evaluation

In this research, first we extract rich high-level features from ECG and Echo images using ViT [39] and Swin Transformer models, respectively. These models capture complementary spatial and structural representations. We then fused the feature vectors extracted from both modalities into a unified, high-dimensional representation per subject. Merging both spatial and structural patterns from multiple modalities results in a feature vector that enhances the understanding of cardiac conditions and improves prediction accuracy. While this fusion enhances the representational strength of the model, it also introduces redundant and irrelevant features, leading to possible overfitting and increased computational complexity. To address this challenge, in this study, we utilize the Aquila Optimizer (AO), a nature-inspired metaheuristic algorithm modeled after the strategic hunting patterns of eagles, to perform feature selection [27]. AO balances global exploration and local exploitation by simulating different stages of eagle hunting, such as soaring at high altitudes to survey large areas and diving swiftly to target specific prey. These adaptive behaviors allow AO to conduct a wide-ranging search across the feature space initially and then gradually refine the search to focus on the most promising feature subsets. This approach effectively reduces the dimensionality of the combined ECG and Echocardiogram feature sets by discarding irrelevant and redundant features, thereby improving classification accuracy and enhancing the generalization performance of the heart disease prediction model. This step was crucial in tailoring the model to focus on medically significant features across both ECG and Echo domains, ultimately enhancing heart disease prediction accuracy.

In this study, to effectively selecting optimal features from the fused ECG and Echocardiogram dataset, we utilize the Aquila Optimizer (AO) [34]. The AO algorithm operates through four strategic movements, each governed by mathematical formulations that simulate exploratory and exploitative behaviors across the feature space [28][37] and the figure6 shows the behavior of Aquila. In the exploration phase**,** AO simulates high soaring with vertical stooping to search broadly across the solution space. The position of each solution (feature subset) is updated using the equation (1):

$$X_{t+1} = X_{mean} - A. | B. X_{mean} - X_t | \quad (1)$$

Here, X_t is the current solution, X_{mean} is the mean position of the population, and A and B are adaptive coefficients controlling exploration range. This formula promotes

global search by encouraging individuals to move away from the current best solution.

During exploitative behavior, the AO simulates contour flight with short glide attacks to fine-tune the search in promising areas. The exploitation mechanism is modeled with formula (2) as:

$$X_{t+1} = G_{best} \cdot \left(1 - \frac{t}{T}\right) + \varepsilon \cdot L \qquad (2)$$

Where G_{best} is the best-so-far solution, t is the current iteration, T is the maximum number of iterations, ϵ is a random coefficient, and L is a Lévy flight step. This phase enhances convergence by gradually reducing the search scope around the global optimum.

To perform binary feature selection, AO solutions are encoded as binary vectors, and a transfer function converts continuous positions into selection probabilities using the equation (3):

$$S(X_{i,j}) = \frac{1}{1+e^{-X_{i,j}}} \qquad (3)$$

A threshold-based rule is then applied:

$$x_{i,j}^{(t+1)} = \begin{cases} 1 & \text{if } \text{rand}() < S(X_{i,j}) \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Where rand () is a uniformly generated random number in the range [0, 1]. This determines whether the j^{th} feature is selected (1) or discarded (0).

The quality of each feature subset is evaluated using a fitness function that balances predictive performance and feature reduction:

$$Fitness_i = \alpha . \text{Accuracy} + \beta . \left(1 - \frac{\text{Number of selected features}}{\text{Total features}}\right) \qquad (5)$$

Here, α and β are weighting factors representing the trade-off between classification accuracy and feature minimization.

By applying these AO-based strategies, the algorithm effectively navigates the feature space, selecting the most informative and non-redundant attributes for accurate heart disease classification. Algorithm 1 presents the datasets acquition, pre-processing, features extraction, fusion and selection.

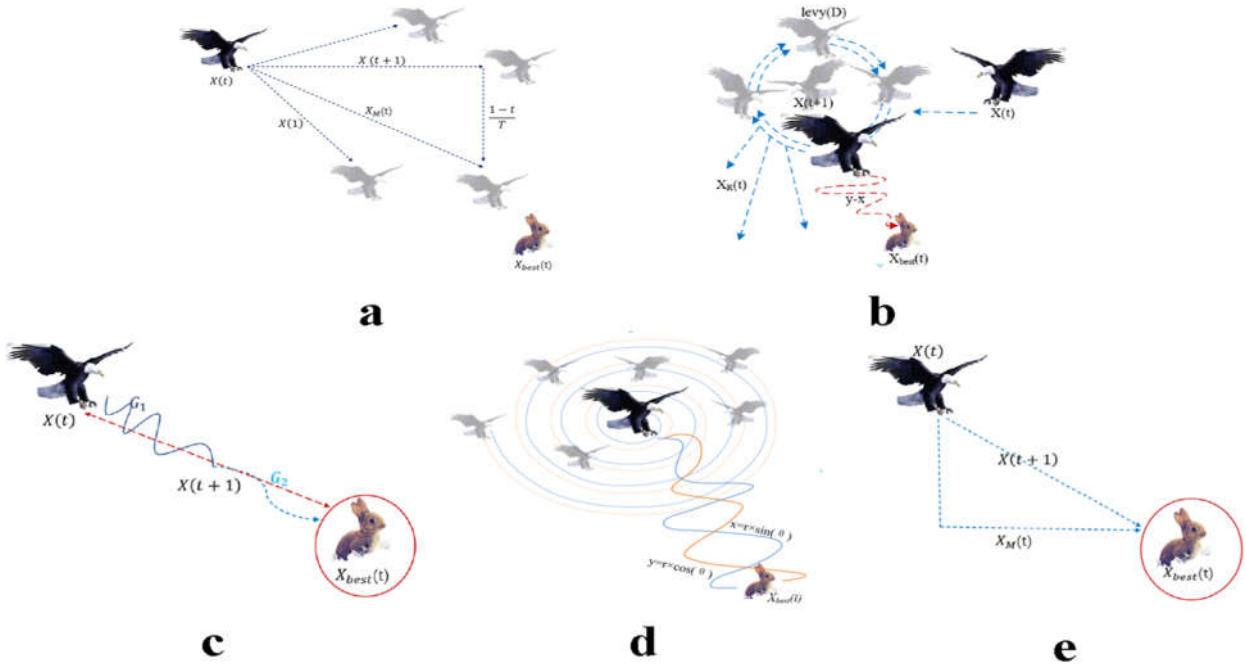

Figure 6: Represents the behavior of Aquila

3.4 Classification Using Fine-Tuned MLP

In our proposed model, we use a Multi-Layer Perceptron (MLP) as the classifier to make the final decision. The MLP uses most relevant features that have been combined and selected from ECG signals and echocardiogram

images. An MLP is a kind of computer brain model with three parts: an entry area, secret areas, and an outcome zone, as depicted in Figures 7 [15] [18]. In every group of brain cells, each cell links to all others in the following group, enabling the system to recognize intricate details within information. In this study, we use one input layer, three hidden layers and one output layer of the MLP classifier design. ReLU acts as the non-linear activator for internal stages, and the sigmoid function operates in the final stage due to our binary categorization goal. We trained the model using the Adam optimizer and the cross-entropy loss function. Important settings such as the learning rate, batch size, number of training rounds (epochs), and the number of hidden layers (3) are carefully adjusted. This setup helped the MLP turn the chosen features into reliable predictions for heart disease. Table 2 shows the specific setup of the MLP classifier, and Algorithm 2 describes how the MLP is designed, trained, and evaluated along with the performance results.

Table 2: Presents used hyperparameters with values of MLP classifier design

No. of hidden layers	3
Input Activation function	ReLu
Optimizer	Adam
Loss function	Cross entropy
Learning rate	0.0001
Batch size	32
Dropout	0.2
No. of epochs	20
Output activation function	Sigmoid

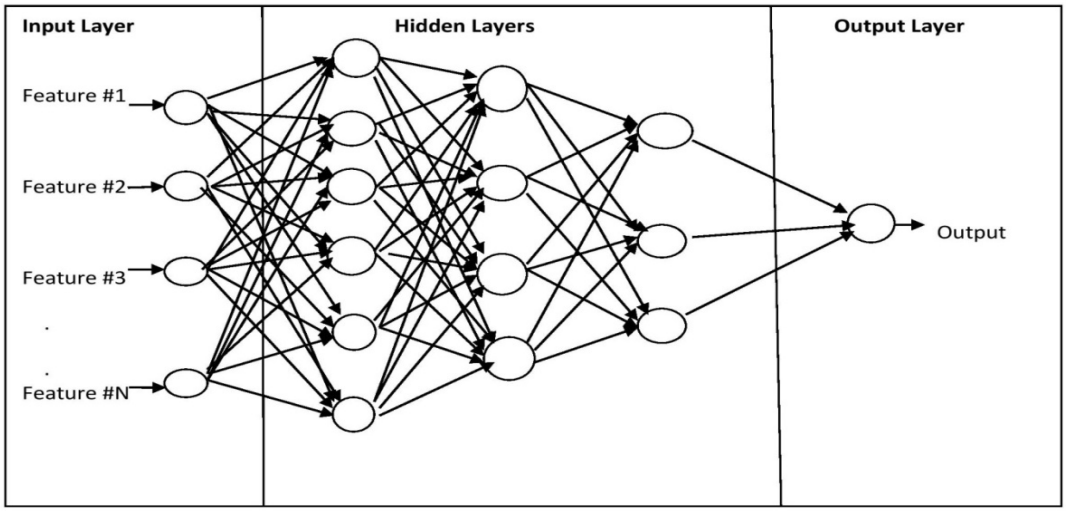

Figure 7: Represents used MLP classifier architecture

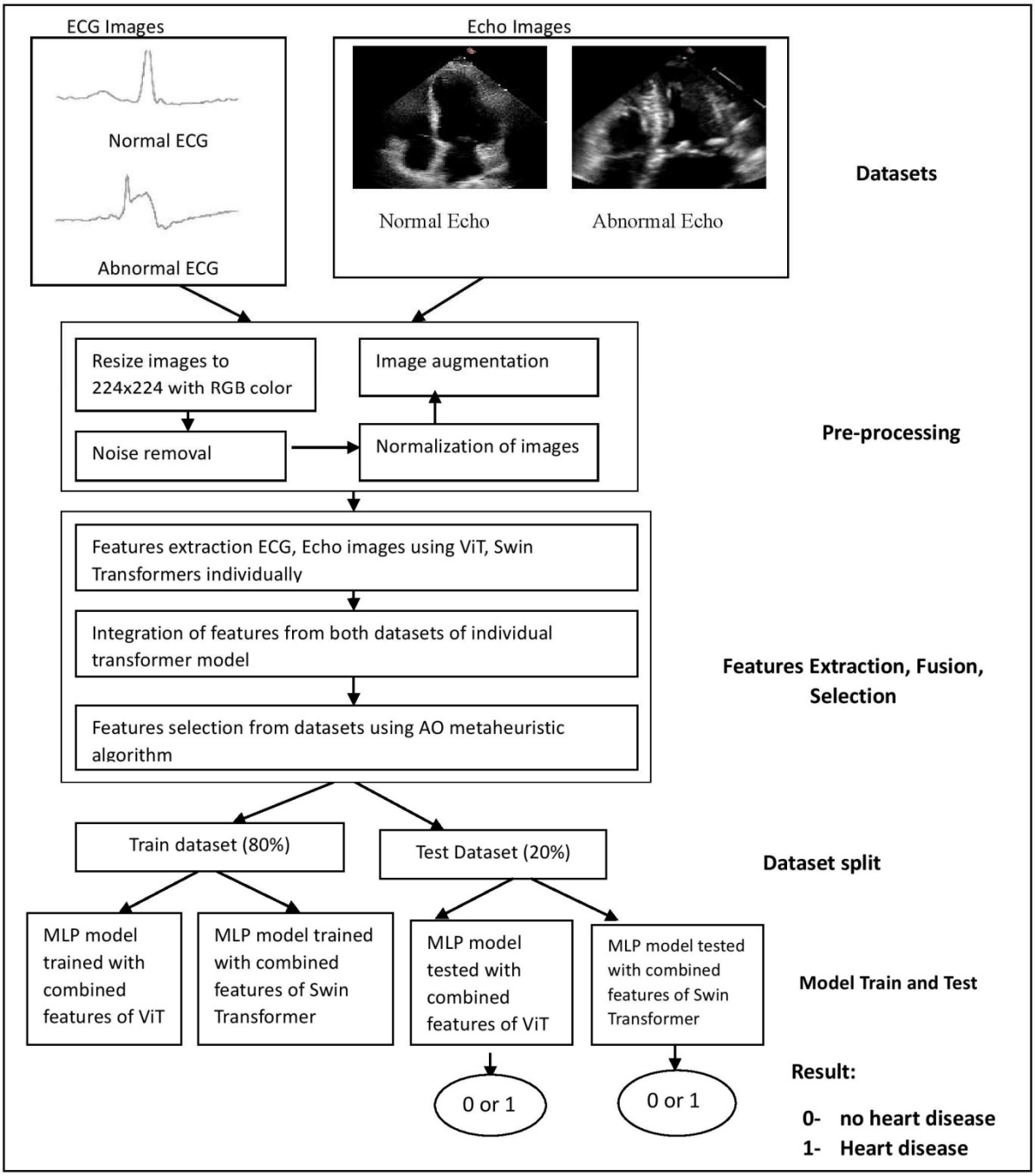

Figure 8: Represents Work flow diagram of Proposed Model

3.5 Performance Evaluation Metrics

To evaluate the efficiency of the proposed deep learning model for heart disease prediction, several performance measures are applied, including accuracy, precision, recall, F1-score, and AUC [12][13]. These indicators provide a comprehensive assessment of how well the model distinguishes between individuals with cardiovascular risk and those without.

Confusion Matrix: The confusion matrix in Figure 9 presents the relationship between the actual and predicted outcomes for heart disease classification, dividing them into positive and negative classes. It is used to calculate accuracy, sensitivity (also called recall), precision, and the F1-score.

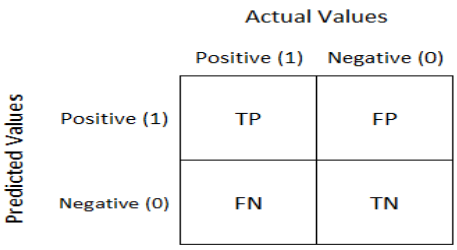

Figure 9: Represents confusion matrix

In this context: TP (True Positive): Correctly predicted positive cases, TN (True Negative): Correctly predicted negative cases, FP (False Positive): Incorrectly predicted as positive, and FN (False Negative): Incorrectly predicted as negative.

Accuracy: Accuracy measures the overall correctness of the model and is expressed as using the equation 5:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision: Precision shows how many of the cases the model predicted as positive are actually positive. It is given by equation 6:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

Recall (Sensitivity): Recall shows how many of the actual positive cases the model correctly identified. Equation 7 is used to calculate recall.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

F1-Score: The F1-score is a way to balance precision and recall by taking their average, which is useful when the classes are not equally common.

$$\text{F1} - \text{score} = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \quad (9)$$

ROC curve: The Receiver Operating Characteristic (ROC) curve demonstrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across different threshold values. These are computed as using equations 10 and 11.

$$\text{TPR} = \frac{TP+FN}{TP} \quad (10)$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad (11)$$

AUC (Area under the Curve): The AUC summarizes the ROC curve into a single value, ranging between 0 and 1. A higher AUC score reflects stronger predictive performance, with values closer to 1 indicating the model's superior ability to separate positive and negative cases effectively.

4. Experimental Results

4.1 Experimental Setup

In this study, we conducted all experiments using Jupyter notebook within the Anaconda3 software environment, utilizing python 3.11.5 as the core programming language with several libraries and frameworks to evaluating our work. Pytorch and TensorFlow are utilized for constructing and training Vision Transformer, Swin Transformer and MLP models. Scikit-learn are utilized for tuning the MLP classifier, evaluating performance metrics, and implementing the AO-PSO fitness function. Numpy and Pandas are utilized for data manipulation and pre-processing, while Opencv and pil are employed for image resizing and transformation. Matplotlib is used to create visualizations like confusion matrices and ROC curves.

4.2 Experimental Results Analysis

Four separate experiments are carried out as shown in table 3, each focusing on a different combination of electrocardiogram (ECG) and echocardiogram (Echo) data. The ECG and Echo images are analyzed using ViT and Swin Transformer models to extract relevant features. These features are then fed into a highly refined MLP classifier. Each configuration is assessed using important performance metrics: accuracy, precision, recall, and f1-score. These experiments aided in comprehending the impact of individual modalities and transformer architectures on the classification task.

Table 3: Represents performance metric values of individual datasets features of ViT, Swin with MLP classifier

Experiment	Accuracy	Precision	Recall	F1-score
Swin-ECG → MLP	86.78%	87.09%	86.78%	86.79%
ViT-ECG → MLP	73.40%	63.05%	73.40%	62.15%
Swin-Echo → MLP	81.70%	79.33%	82.66%	80.66%
ViT-Echo → MLP	73.37%	74.33%	73.37%	73.64%

In this study,we employs multimodal fusion by combining features extracted individually from ECG and Echo images using Vision Transformer (ViT) and Swin Transformer architectures. For each transformer, ECG and Echo features are first extracted independently and then fused to form a comprehensive feature representation. To refine this high-dimensional fused vector and eliminate redundant or irrelevant attributes, we introduce an Aquila Optimizer for feature selection. This method effectively enhanced model generalization while reducing complexity. The optimized feature subsets are then classified using a fine-tuned Multi-Layer Perceptron (MLP) model. Results indicate that the Swin Transformer-based fusion, optimized with AO and classified using MLP, achieved the best performance (96.76% accuracy and 97.98% F1-score), surpassing the ViT-based fusion (87.68% accuracy and 92.75% F1-score). These findings validate the effectiveness of combining multimodal features with optimization and highlight the superior discriminative power of Swin Transformer in capturing cardiac image patterns for accurate heart disease prediction.This performance metric result is outlined in table 4.

Table 4: Presents performance metric values of combined features of ViT with MLP and combined features of Swin with MLP classifier

Experiment	Accuracy	Precision	Recall	F1-score
Swin-ECG+Swin-Echo+ MLP	96.76%	96.19%	99.84%	97.98%
ViT-ECG + ViT-Echo + MLP	87.68%	86.48%	100%	92.75%

The experimental findings clearly show a significant difference in performance between single-modality and multimodal fusion approaches in heart disease prediction. Among the single-modality models, Swin Transformer -ECG → MLP showed better accuracy (86.78%) than other configurations, especially outperforming ViT-ECG → MLP (73.40%). Similarly, Swin Transformer -Echo → MLP (81.70%) slightly surpassed ViT-Echo → MLP (73.37%). However, the most significant performance boost is observed in the multimodal fusion. The combination of Swin Transformer-extracted ECG and Echo features, optimized using the AO algorithm and classified with a fine-tuned MLP, achieved the highest accuracy of 96.76% and an F1-score of 97.98%. The Swin-based fusion also improved over individual models but remained lower, with 87.68% accuracy and a 92.75% F1-score. These outcomes highlight the effectiveness of feature fusion and the strength of AO in selecting informative features for enhanced prediction. Our experimental findings reveal that although Vision Transformer (ViT) demonstrates good performance in heart disease prediction, Swin Transformer delivers superior results, especially in terms of accuracy and feature extraction. The hierarchical structure of Swin Transformer enables it to learn more effectively from both ECG and Echo images, leading to improved predictive performance.

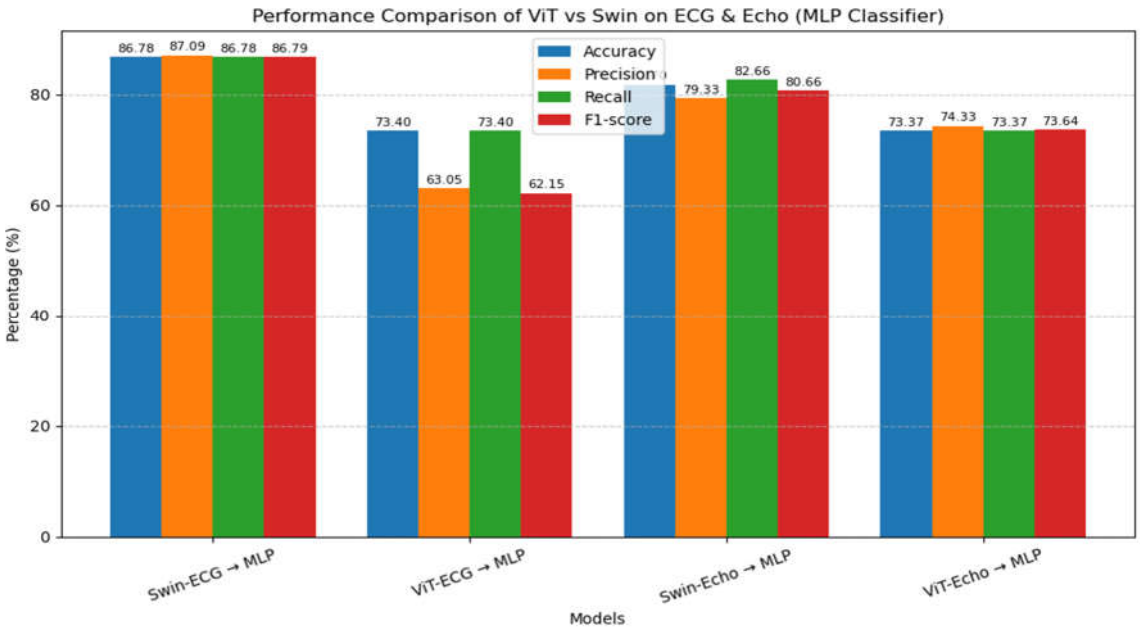

Figure 10: Represents Performance Comparison of ViT and Swin with MLP on individual datasets

4.3 Model Performance Metrics Analysis

To evaluate the effectiveness of the proposed multimodal fusion approaches for heart disease prediction, we analyzed the confusion matrices of both models, along with key performance metrics including accuracy, precision, recall, and f1-score. We also evaluated the ROC-AUCscores. These assessments are carried out for both fusion pipelines: ViT-ECG + ViT-Echo + MLP and Swin-ECG + Swin-Echo + MLP.

Confusion Matrix Analysis: Confusion matrices are plotted for the ViT-ECG + ViT-Echo + MLP and Swin-ECG + Swin-Echo + MLP models to visually assess prediction outcomes in terms of true positives, true negatives, false positives, and false negatives. The model's confusion matrix showcased a significant presence along the diagonal, suggesting exceptional accuracy in classifying data with minimal errors. In contrast, the Swin-based model, while somewhat effective, exhibited a few more off-diagonal elements, indicating a higher frequency of incorrect predictions.

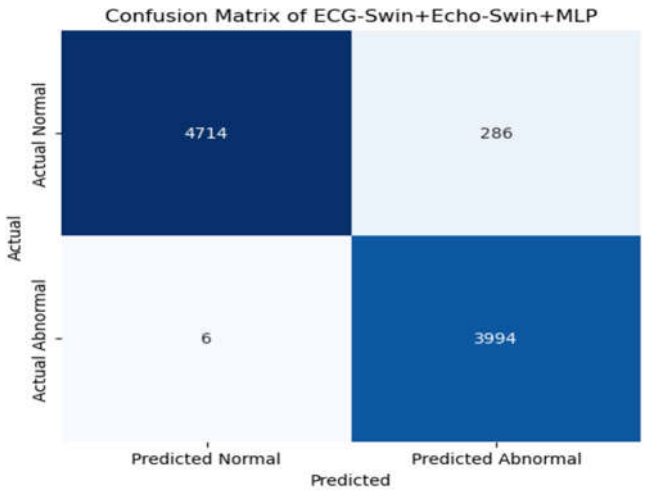

Figure 11: Represents confusion matrix of ECG-Swin+Echo-Swin+MLP

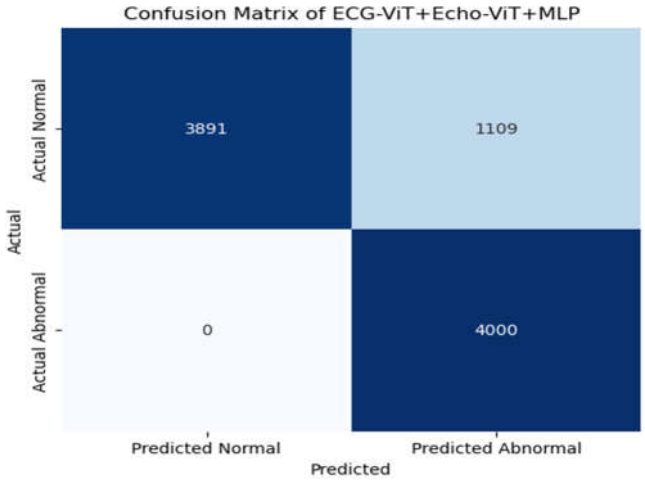

Figure 12: Represents confusion matrix of ECG-ViT+Echo-ViT+MLP

Accuracy, Recall, Precision and F1 score Analysis: Performance metrics including accuracy, precision, recall, and f1-score are evaluated for both fusion models. The Swin Transformer-ECG + Swin Transformer-Echo + MLP configuration outperformed with an accuracy of 96.76%, precision of 96.19%, recall of 99.84%, and an f1-score of 97.98% as shown in figure 13 . Meanwhile, the ViT-based model reached 87.68% accuracy, 86.48% precision, perfect recall of 100%, and a 92.75% f1-score as presented in figure 14.These results highlight the Swin Transformer-based fusion model's advantage in delivering more balanced and robust predictive performance.

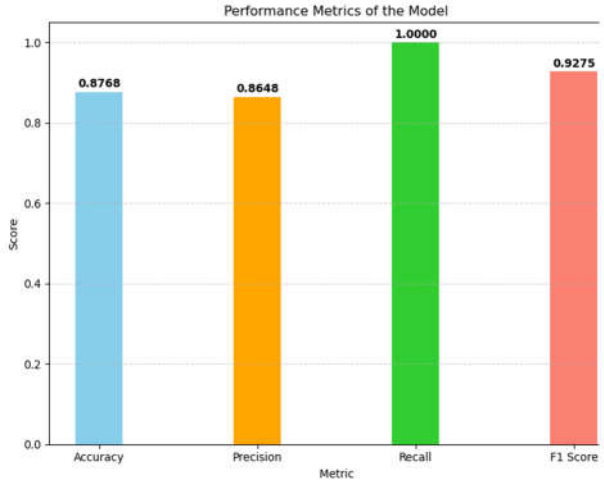

Figure 13: Performance Metrics of ViT-ECG + ViT-Echo + MLP

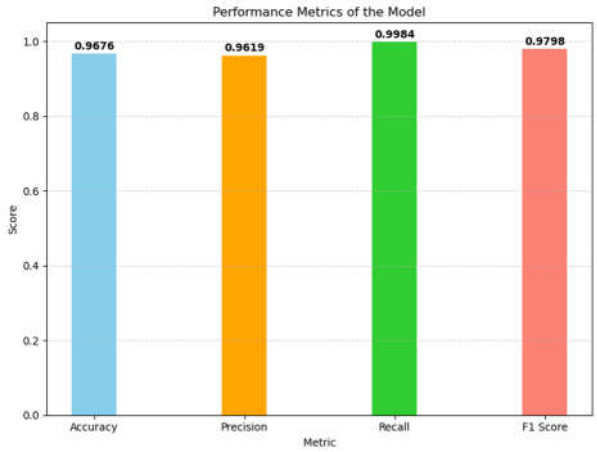

Figure 14: Performance metrics of Swin-ECG+Swin-Echo+ MLP

ROC and AUC Analysis: Furthermore, we plotted ROC curves for both fusion models to evaluate their ability to differentiate between the presence and absence of heart disease at different thresholds. The areaunder the curve (AUC) for the ViT-based fusion model was closer to 1, indicating a high level of discrimination. The Swin-based fusion model also achieved a high AUC, but it was slightly lower than that of the ViT-based model. The results of these visually presented in figure 15and figure 16.

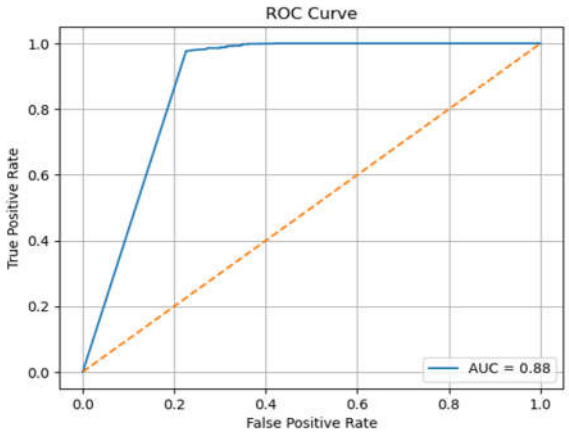

Figure 15: Results of ROC curve with AUC value of ViT-ECG+ViT-Echo+ MLP

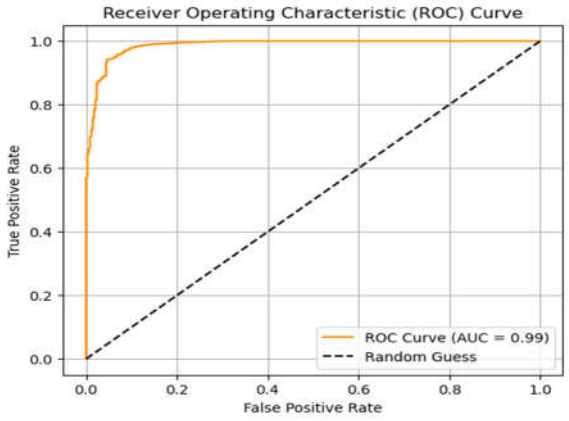

Figure 16: Results of ROC curve with AUC of Swin-ECG+Swin-Echo+ MLP

4.4 Comparative Performance

We compare the performance of our proposed a Two-Stage Hybrid Deep Learning Model with already existed single and multimodality models. The table 2 presents a comparative

evaluation of heart disease prediction models based on accuracy reported by various researchers and also figure 17 shows the accuracy of highlighting the superior performance of our model.

Table 5: Comparison of different existing models with our proposed model

Author name, Reference No	Model used	Accuracy
N. Revathi et al.,[20]	MLP	75.60%
Tb Ai Munandar [21]	MLP with Tanh	78.80%
Songhee Cheon et al.,[26]	PCA with DNN	83.48%
Majumder et al.,[36]	Random Forest with WOA	86. 53%
Pengpai Li et al.,[22]	LSTMs+GA	87.3%
Archana Agarwal et al.,[33]	Random Forest	90.24%
Mohammad Shokouhifar et al.,[30]	EHMFFL	91.8%
Geetha Narasimhan et al.,[24]	GAORF	92%
Prabu Pachiyannan et al.,[14]	ML-CHDPM(LSTM+Attention Mechanism)	94.28%
Muhammad Tayyeb et al.,[15]	MLP	94.40%
Muhammad Mahtab et al.,[3]	EfficientNetB3	94.53%
Fande Kong et al.,[23]	LDGO	94. 6%
Muhammad Raoof et al.,[2]	VGG16	94.92%
Zeynep Hilal Kilimci et al.,[5]	BeiT	95.90%
Rayudu Srinivas et al.,[40]	ACLS-RCNN model with ICSOA	95.64%
Proposed Model	Two-Stage Hybrid Deep Learning Model	96.76%

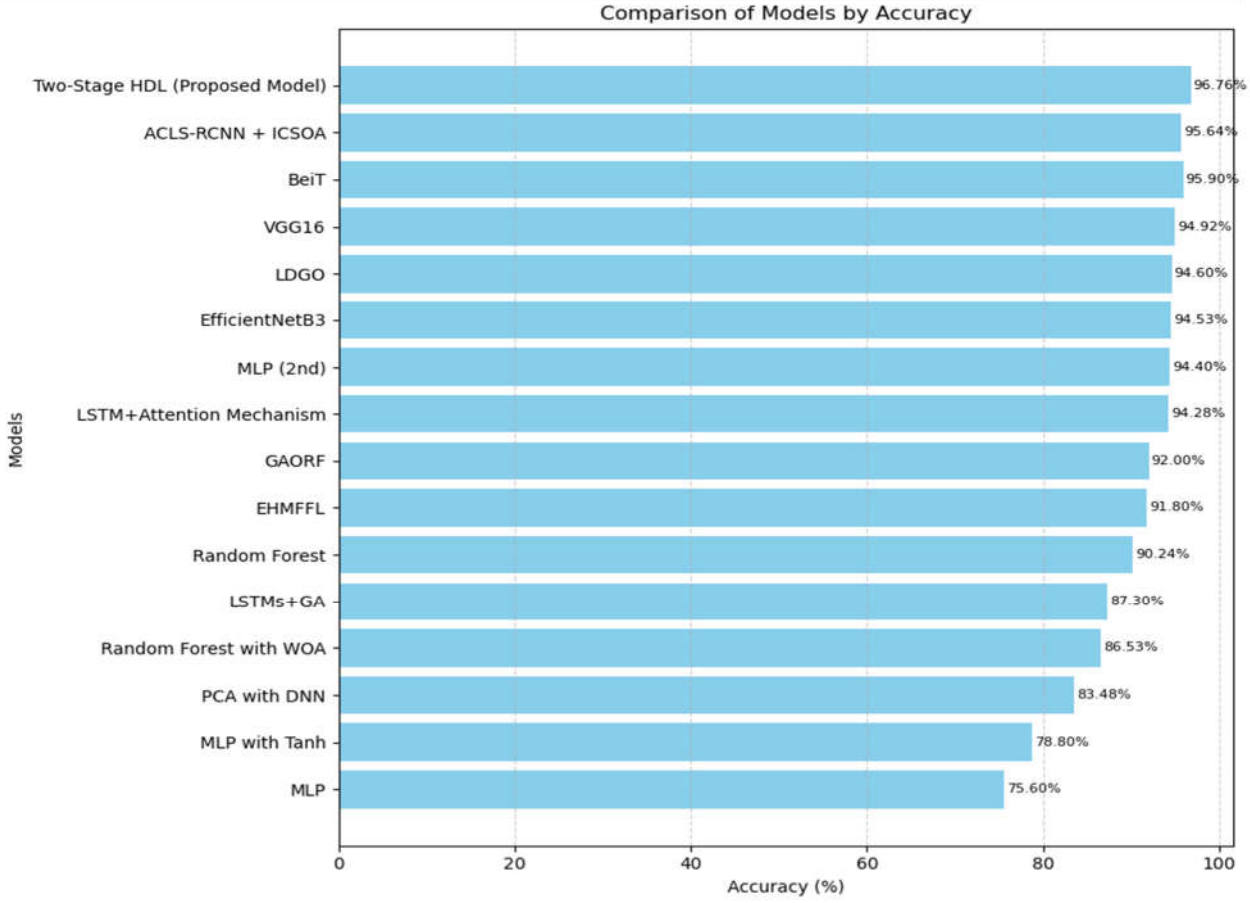

Figure 17: Represents accuracy comparison of proposed model with existing models

5. Conclusion and Future work

In this study, we introduce an advanced deep learning framework for heart disease prediction using ECG and Echocardiogram (Echo) images. Feature extraction was performed using Vision Transformer (ViT) and Swin Transformer architectures, followed by classification through a fine-tuned Multi-Layer Perceptron (MLP). Among all configurations, the fusion of ECG and Echo features extracted via Swin Transformer and optimized using an Aquila Optimization algorithm demonstrated the best performance, achieving 96.76% accuracy and a 97.98% F1-score. The integration of AO effectively refined the fused feature set by selecting the most relevant features, enhancing classification accuracy while reducing complexity. Findings confirm that integrating multimodal fusion with transformer-based extraction and metaheuristic optimization substantially enhances predictive outcomes. This study highlights that Swin Transformer, when utilized with ECG and Echo image data, achieves better predictive accuracy, more efficient feature extraction, and enhanced generalization compared to Vision Transformer (ViT). The hierarchical attention mechanism employed by Swin Transformer allows it to better capture both local and global features, thereby making it a more effective model for multimodal heart disease prediction. Future directions include incorporating clinical tabular data, applying attention-based fusion techniques, tuning hyperparameters, expanding the dataset, and validating the model in clinical environments to further increase reliability and generalizability.

References

1. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Muhammad Raoof, Muhammad Mahtab, SohailMasoodBhatti, Muhammad Rashid and ArfanJaffar, “Heart Disease Classification From Echocardiogram Images Using Deep Learning”, IEEE Access, 14 January 2025,Volume 13, 2025, doi:10.1109/ACCESS.2024.3524732.
3. Muhammad Mahtab, ZainabSadiq, Muhammad Raoof, and SohailMasoodBhatti, “ Enhancing Heart Disease Detection in Echocardiogram Images Using Optimized EfficientNetB3 Architecture”, Journal of Computing & Biomedical Informatics , Volume 07 Issue 02 2024, September 01, 2024, https://doi.org/10.56979/702/2024, ISSN: 2710- 1606.
4. Chunyu Fan, Ql Su, Zhifeng Xiao, Hao Su, AijieHou, and Bo Luan, “ ViT-FRD: A Vision Transformer Model for Cardiac MRI Image Segmentation Based on Feature Recombination Distillation”, IEEE Engineering in Medicine and Biology Society Section, IEEE Access, 22 November 2023, Volume 11,2023, doi: 10.1109/ACCESS.2023.330222.
5. Zeynep Hilal Kilimci, MusatafaYalcin, AyhanKucukmanisa and Amit Kumar Mishra, “ Heart Disease Detection using Vision-Based Transformer Models from ECG Images”,
6. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, XiaohuaZhai, Thomas Unterthiner, MostafaDehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, JakobUszkoreit, and Neil Houlsby, “ An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, ICLR 2021, 3 June 2021.
7. Mahmoud Khalil, Ahmad Khalil and Alioune Ngom , “ A Comprehensive Study of Vision Transformers in Image Classification Tasks”, 5 Dec 2023.
8. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo,”Swin Transformer: Hierarchical

Vision Transformer using Shifted Windows", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), *Aug 2021,* pp:10012-10022, https://doi.org/10.48550/arXiv.2103.14030.

9. Koki Nakanishi, Shunichi Homma, "Role of echocardiography in patients with stroke", Journal of Cardiology, 30 May 2016, http://dx.doi.org/10.1016/j.jjcc.2016.05.001.
10. Priya Dubey, Dipti RanjanTiwari, " Heart disease Prediction using Deep Learning Techniques: A Review", IRJET, e-ISSN:2395-0056, Volume: 11 Issue:09, Sep 2024,www.irjet.net
11. Lerina Aversano, Mario Luca Bernardi, Marta Cimitile, Debora Montano, and Riccardo Pecori, " Early Diagnosis of Cardiac Diseases using ECG Images and CNN-2D", Procedia Computer Science 225(2023) 2866-2875, www.sciencedirect.com, doi:10.1016/j.procs.2023.10.279.
12. G. Sunil kumar and P. Kumaresan, "Deep Learning and Transfer Learning in Cardiology: A Review of Cardiovascular Disease Prediction Models", IEEE Access, Volume 12, December 2024, pp: 193365-193386, DOI:10.1109/ACCESS.2024.3514093.
13. Mohammed Amine Bouqentar, OumaimaTerrada, SoufianeHamida, ShawkiSaleh, DrissLamrani, BouchaibCherradi and AbdelhadiRaihani, "Early heart disease prediction using feature engineering and machine learning algorithms", www.cell.com/heliyon, October 2024,2405-8440, https://doi.org/10.1016/j.heliyon.2024.e38731.
14. Prabu Pachiyannan, Musleh Alsulami, Deafallah Alsadie, Abdul Khader Jilani Saudagar, Mohammed AlKhathami and Ramesh Chandra Poonia, "A Novel Machine Learning-Based Prediction Method for Early Detection and Diagnosis of Congenital Heart Disease Using ECG Signal Processing",Technologies **2024**, 12, 4. https://doi.org/10.3390/technologies12010004.
15. Muhammad Tayyeb, MuhammadUmer, KhaledAlnowaiser, SaimaSadiq, Ala' AbdulmajidEshmawi, RizwanMajeed, AbdullahMohamed, Houbing Song and Imran Ashraf, "Deep Learning Approach for Automatic Cardiovascular Disease Prediction Employing ECG Signals", Computer Modeling in Engineering & Sciences (CMES), June 2023, vol.137, no.2, pp: 1678-1694,DOI: 10.32604/cmes.2023.026535.
16. Priti Shinde, Mahesh Sanghavi, TienAnh Tran, "A Survey on Machine Learning Techniques for Heart Disease Prediction", SN Computer Science (2025) 6:334, April 2025, https://doi.org/10.1007/s42979-025-03860-2.
17. Atta Ur Rahman, YousefAlsenani, AdeelZafar, KalimUllah, KhaledRabie and ThokozaniShongwe, "Enhancing heart disease prediction using a self attention based transformer model", www.nature.com/scientificreports, 2024 14:514, https://doi.org/10.1038/s41598-024-51184-7.
18. R. Saravana Ram, J. Akilandeswari and M. Vinoth Kumar, " HybDeepNet: A Hybrid Deep Learning Model for Detecting Cardiac Arrhythmia from ECG Signals", Information Technology and Control, Vol. 52 / No. 2 / 2023 ,pp. 433-444 ,DOI 10.5755/j01.itc.52.2.32993.
19. Khalid Al-hammuri, Fayez Gebali, AwosKanan and Ilamparithi Thirumarai Chelvan,"Vision transformer architecture and

applications in digital health: a tutorial and survey", Visual Computing for Industry, Biomedicine, and Art (2023) 6:14, https://doi.org/10.1186/s42492-023-00140-9.

20. N. Revathi, P.M. Kavitha, D. Narayani, S.Irin Sherly M. Robinson Joel and P.Jose, "Heart Disease Prediction using Multimodal Data with Multi-Layer Perceptron", www.ijisae.org, 2024, 12(4),pp.1728-1737.
21. Tb Ai Munandar, " Data Mining for Heart Disease Prediction Based on Echocardiogram and Electrocardiogram Data", JOIN(Jurnal Online Informatika), e-ISSN:2527 – 9165, Volume 8 Number 1, June 2023, doi: 10.15575/join.v8i1.1027.
22. Pengpai Li, Yongmei Hu, Zhi-Ping Liu, " Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods", Biomedical Signal Processing and Control, www.elsevier.com/locate/bspc, Feb 2021, https://doi.org/10.1016/j.bspc.2021.102474.
23. Fande Kong, Zhengyi Song, and Qijia Liu, "The frontiers of intelligent health services: cardiovascular disease prediction using novel machine learning methods and metaheuristic algorithm", COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING, May 2025, www.tandfonline.com/journals/gcmb20, https://doi.org/10.1080/10255842.2025.2502823.
24. Geetha Narasimhan & Akila Victor, "A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction", www.nature.com/scientificreports, (2025) 15:10971, https://doi.org/10.1038/s41598-024-73867-x.
25. Jingyuan Yi, Peiyang Yu, Tianyi Huang, and ZeqiuXu, " Optimization of Transformer Heart Disease Prediction Model based on Particle Swarm Optimization Algorithm", 7 Jan 2025.
26. Songhee Cheon , Jungyoon Kim and Jihye Lim ," The Use of Deep Learning to Predict Stroke Patient Mortality", International Journal of Environmental Research and Public Health, May 2019, www.mdpi.com/journal/ijerph, doi:10.3390/ijerph16111876.
27. Suqian Wu , Bitao He , Jing Zhang , Changshen Chen and Jing Yang, " PSAO: An enhanced Aquila Optimizer with particle swarm mechanism for engineering design and UAV path planning problems", www.elsevier.com/locate/aej, August 2024, https://doi.org/10.1016/j.aej.2024.08.021.
28. Buddhadev Sasmal , Arunita Das , Krishna Gopal Dhal and Swarnajit Ray," Aquila-particle swarm based cooperative search optimizer with superpixel techniques for epithelial layer segmentation", Applied Soft Computing Journal, www.elsevier.com/locate/asoc,October2023, https://doi.org/10.1016/j.asoc.2023.110947.
29. SUSHILA PALIWAL, SURAIYA PRAVEEN, M. AFSHAR ALAM, JAWED AHMED," OPTIMIZING HEART DISEASE PREDICTION MODELS USING GENETIC ALGORITHMS: A METAHEURISTIC APPROACH", Journal of Theoretical and Applied Information Technology, May 2024.

Vol.102. No 9, ISSN: 1992-8645, www.jatit.org,

30. Mohammad Shokouhifar , Mohamad Hasanvand, Elaheh Moharamkhani and Frank Werner," Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data", Algorithms 2024, 17, 34. https://doi.org/10.3390/a17010034, January 2024, https://www.mdpi.com/journal/algorithms.
31. Ashish Vaswani,Noam Shazeer,Niki Parmar,Jakob Uszkoreit,Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need", 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
32. Parul Madan, Vijay Singh, Devesh Pratap Singh, Manoj Diwakar, Bhaskar Pant and Avadh Kishor, "A Hybrid Deep Learning Approach for ECG-Based Arrhythmia Classification", Bioengineering 2022, 9, 152. https://doi.org/10.3390/bioengineering9040152, April 2022, https://www.mdpi.com/journal/bioengineering.
33. Archana Agarwal, Ravi Singh Bajetha, Saurav Dhyani, Anmol Pal, Akshit Bhatt, " Heart Disease Prediction Using Machine Learning", International Journal of Creative Research Thoughts (IJCRT) , Volume 12, Issue 5 May 2024 , ISSN: 2320-2882, www.ijcrt.org.
34. Himanshu Sharma, Krishan Arora, Raghav Mahajan, Syed Immamul Ansarullah, Farhan Amin & Hussain AlSalman, "Improved aquila optimizer for swarm-based solutions to complex engineering problems", www.nature.com/scientificreports, December 2024, https://doi.org/10.1038/s41598-024-79577-8.
35. Government General Hospital, Cardiology Department. (n.d.)(2025*).* Patient data collected from Ananthapuramu, Andhra Pradesh. Government General Hospital.
36. Annwesha Banerjee Majumder, Somsubhra Gupta, Dharmpal Singh,Sourav Majumder, "An Advanced Model to Predict Heart Disease Applying Random Forest Classifier and Whale Optimization Algorithm", Indian Journal of Science and Technology 2023;16(43):3679–3690, 11-11-2023, https://www.indjst.org/, https://doi.org/10.17485/IJST/v16i41.1756,
37. Youfa Fu , Dan Liu, Shengwei Fu , Jiadui Chen & Ling He, "Enhanced Aquila optimizer based on tent chaotic mapping and new rules", www.nature.com/scientificreports, (2024) 14:3013, https://doi.org/10.1038/s41598-024-53064-6.
38. Qiumei Pu, Zuoxin Xi, Shuai Yin, Zhe Zhao and Lina Zhao, "Advantages of transformer and its application for medical image segmentation: a survey", *BioMedical Engineering OnLine (2024) 23:14,* https://doi.org/10.1186/s12938-024-01212-4.
39. Ch.Sita Kameswari, Kavitha J, T. Srinivas Reddy, Balaswamy Chinthaguntla, Senthil Jagatheesaperumal, Silvia Gaftandzhieva, Rositsa Doneva, "An Overview of Vision Transformers for Image Processing: A Survey", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14, No. 8, January 2023,* www.ijacsa.thesai.org.
40. Rayudu Srinivas, Ravi kiran Bagadi, T. Rama Reddy, Neti Praveen, G.

Aparanjini, “An efficient hybrid optimization algorithm for detecting heart disease using adaptive stacked residual convolutional neural networks”, Biomedical Signal Processing and Control 87 (2024) 105522, www.elsevier.com/locate/bspc, October 2023, https://doi.org/10.1016/j.bspc.2023.105522.