SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 7012-7046
Open Access. Online & Print

DOI: 10.5281/zenodo.12426719

AQUILA-GUIDED CONTEXT ADAPTIVE MULTIMODAL INCEPTION NETWORK FOR STRONG RHEUMATIC HEART DISEASE CLASSIFICATION

Jagadesh A N[1*], Ravikumar M[2], Indrakumar K[3]

[1,3] Research Scholar, Department of Computer Science, Jnanasahyadri, Kuvempu University, Shivamogga 577 451, INDIA.
[2]Professor, Department of Computer Science, Jnanasahyadri, Kuvempu University, Shivamogga 577 451, INDIA.

Received: 04/08/2025
Accepted: 07/01/2026

Corresponding author: Jagadesh A N
(jagaenator@gmail.com)

ABSTRACT

Rheumatic Heart Disease (RHD) is a major cardiovascular issue globally, especially in low- and middle-income countries where early diagnosis is often limited. While deep learning shows promise in automated cardiac diagnosis, current methods often struggle to effectively combine different cardiac modalities, model their relationships, and optimize complex multimodal structures. To address these challenges, this study introduces an Aquila-Guided Context-Adaptive Multimodal Inception Network designed to improve RHD classification using echocardiographic images and phonocardiographic signals. [1]. The framework includes multiscale Inception-based feature extraction and a channel attention mechanism that adjusts features based on context. It also features a cross-modal attention fusion module that combines structural and acoustic cardiac representations. To improve optimization, we use a bilevel hyperparameter tuning framework. In this setup, network weights are trained with gradient-based learning, while hyperparameters are optimized globally using the Aquila Optimization Algorithm (AO). [7]. Additionally, the model uses focal loss to address class imbalance and applies Youden Index-based threshold tuning together with temperature scaling for probability calibration to boost diagnostic confidence. Experimental results show that the proposed framework outperforms several baseline models, including CNN, ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer. It achieves an accuracy of 0.91 and an AUC of 0.91, demonstrating strong sensitivity and specificity for clinical screening. Comprehensive validation through ablation studies, cross-validation, optimizer comparisons, statistical tests, and explainability analysis confirms the reliability and transparency of this approach. These findings suggest that combining adaptive multimodal fusion with global hyperparameter tuning can greatly improve automated detection of Rheumatic Heart Disease and support AI-assisted cardiac screening in healthcare settings with limited resources. [22][29]

KEYWORDS: Natural Rheumatic Heart Disease, multimodal deep learning, echocardiography analysis, phonocardiogram classification, cross-modal attention fusion, Aquila optimization algorithm, medical image analysis, and explainable artificial intelligence (XAI

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License. (https://cre-ativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

1.1 Clinical Background and Motivation

Rheumatic Heart Disease (RHD) remains a significant global health problem [1], especially in low- and middle-income countries where access to early diagnostic tools is limited. The World Health Organization estimates that more than 39 million people worldwide [3] are affected by RHD. It continues to be a leading cause [4] of heart issues and premature death in developing regions. RHD primarily develops from untreated Group A streptococcal infections, which cause acute rheumatic fever and persistent inflammation that damages the heart valves [2]. This damage can lead to valve stenosis, regurgitation, heart failure, and other severe complications that greatly diminish patients' quality of life. Recognizing RHD symptoms early is crucial for prompt treatment and preventing permanent heart damage.

Despite its importance, early detection of RHD remains challenging due to various clinical and infrastructure-related barriers. Doctors typically rely on echocardiography, auscultation, and physical exams for diagnosis. Echocardiography is considered the gold standard because it provides detailed images of the heart's valves and chambers, including their structure and function [7][8]. However, echocardiographic screening requires specialized equipment and trained cardiologists, making it difficult to implement in areas with limited resources or rural settings where RHD is most prevalent [10]. Additionally, the accuracy of echocardiographic results depends heavily on the operator, and diagnostic precision can vary with the clinician's experience.[11]

Another challenge in automated cardiac diagnostics arises from the different types of clinical data. Echocardiographic images show the structure of the heart valves and chambers, while phonocardiographic recordings capture sounds associated with turbulent blood flow and valve abnormalities. Each type of data provides valuable information, but combining these sources is complicated [22]. Many existing models rely on only one type of data or use static methods that do not fully utilize the relationships between data types. This limits the accuracy of diagnoses.

These challenges drive the development of innovative multimodal diagnostic systems that combine various cardiac data and highlight clinically important features. By integrating echocardiographic imaging with phonocardiographic signal analysis and using advanced deep learning techniques, it becomes possible to develop automated systems that improve diagnostic accuracy, reduce reliance on specialists, and enable large-scale screening in healthcare settings with limited resources.

1.2 Limitations of Existing Deep Learning Approaches

Despite rapid progress in deep learning for medical image and signal analysis, several challenges remain in the automatic detection of Rheumatic Heart Disease (RHD) [12]. Current deep learning systems often struggle to effectively integrate diverse cardiac data, ensure consistency across patient groups, and remain reliable in real-world clinical settings [22]. These problems stem from architectural, optimization, and data limitations that impact the reliability of current models.

A major limitation is the reliance on static multimodal fusion techniques [25]. Many existing multimodal diagnostic systems combine features from different modalities using simple concatenation or fixed-weighting methods. These methods assume that each modality contributes equally to the diagnostic outcome. However, in clinical practice, the reliability of data sources can vary significantly due to noise, motion artifacts, signal distortion, or differences in imaging equipment. Static fusion techniques cannot adjust the importance of each modality based on data quality, which reduces the robustness of automated diagnostic systems.[12]

Another limitation is the weak modeling of cross-modal dependencies. In multimodal cardiac diagnosis, echocardiographic images reveal the structure of cardiac valves and chambers, while phonocardiographic recordings provide acoustic information about blood flow and valve abnormalities. These modalities are physiologically connected. Yet, many deep learning models treat them as separate feature streams before combining them later in the network [22]. This underuses important interactions between structural and acoustic patterns, limiting the model's ability to identify meaningful clinical relationships.

An additional challenge comes from over-reliance on gradient-based local optimization methods in model training. Most deep learning frameworks use optimizers such as stochastic gradient descent (SGD) or Adam to update the network parameters [27]. While these methods are efficient, they mainly perform local searches in high-dimensional parameter spaces. In complex architectures with attention mechanisms, multimodal fusion layers, and many hyperparameters, the optimization landscape becomes highly non-convex. As a result, gradient-based optimizers may become trapped in suboptimal local minima, leading to unstable or poor model

performance.

Furthermore, many existing deep learning models suffer from poor calibration and sensitivity to class imbalance, which are common in medical datasets [29] . RHD screening datasets often contain significantly fewer pathological cases compared to normal samples. Standard training methods using categorical cross-entropy loss tend to bias the model toward the majority class, lowering its sensitivity to clinically important minority cases. Additionally, deep neural networks often produce poorly calibrated probability estimates, meaning their predicted confidence scores do not accurately reflect the actual likelihood of outcomes [34]. This lack of calibration can harm clinical trust and impede the practical deployment of AI-based diagnostic systems.

These limitations highlight the need for more advanced deep learning frameworks that can dynamically combine multimodal cardiac data, understand relationships among different data types, and enable more reliable optimization and calibration. Addressing these issues is essential for creating trustworthy AI systems that allow early and accurate detection of Rheumatic Heart Disease in real-world healthcare settings.

1.3 Research Gap

The limitations discussed earlier highlight several ongoing challenges in developing reliable deep learning systems for detecting Rheumatic Heart Disease (RHD). Although deep neural networks have enhanced the analysis of medical images and physiological signals, current methods still struggle to manage multimodal cardiac data, optimize complex model parameters, and provide clear diagnostic explanations [22]. Solving these issues is crucial for creating AI systems suitable for clinical use.

A significant gap exists because current multimodal learning systems lack adaptive modality weighting [24]. Most models use fixed fusion methods, in which features from echocardiographic images and phonocardiographic signals are combined via static concatenation or with fixed weights. However, in real clinical settings, the reliability of each modality can vary with acquisition conditions, signal quality, and patient characteristics. Without the ability to adjust the importance of each modality, models risk relying too much on noisy or low-quality inputs, which can hurt classification accuracy. Therefore, there is a clear need for adaptive fusion techniques that automatically adjust the contributions of modalities based on contextual relevance.

Another notable gap is the lack of global optimization [27] methods for multimodal hyperparameters. Deep learning models with attention modules, fusion layers, and other components often have many hyperparameters. Traditional training methods primarily rely on gradient-based optimizers such as Adam or stochastic gradient descent. While these methods effectively tune network weights, they are less efficient for exploring the complex hyperparameter space of multimodal architectures. Consequently, these approaches may lead to suboptimal configurations, thereby limiting the model's overall performance and stability. Implementing global optimization techniques that balance exploration and exploitation can help identify better hyperparameter settings and enhance the reliability of the learning process.

Another research gap is the lack of clinically interpretable deep learning models. In medical decision-making, model transparency is essential for building trust among clinicians and healthcare providers. Many deep learning architectures act as "black boxes," making predictions without clearly explaining their decision processes. This lack of interpretability [34][35] can impede clinical adoption, especially for high-risk tasks such as detecting cardiac disease. Therefore, there is a growing need for models that not only achieve high classification accuracy but also offer interpretable visualizations, such as attention maps or activation heatmaps, that highlight diagnostically relevant areas in medical data.

These research gaps highlight the need for a diagnostic framework that integrates adaptive multimodal fusion, global hyperparameter optimization, and explainable learning techniques. Tackling these issues can greatly improve the reliability, robustness, and clinical application of AI systems for the early detection of Rheumatic Heart Disease.

1.4 Contributions

To overcome the limitations of current multimodal diagnostic systems and address the research gaps mentioned earlier, this study introduces a deep-learning framework for automatically classifying Rheumatic Heart Disease (RHD) symptoms. The model integrates multimodal cardiac data with enhanced attention mechanisms and training strategies to boost diagnostic accuracy and clinical clarity. The main contributions of this work are summarized below:

1. Context-Adaptive Cross-Modal Attention Fusion:

This new multimodal fusion method adjusts the

importance of various modalities based on context, enabling better modeling of structural and acoustic cardiac connections.

2. Bilevel Hyperparameter Optimization with Aquila Algorithm:

This global optimization framework simultaneously tunes network weights and hyperparameters. It enhances model stability and prevents getting stuck in local minima in complex multimodal configurations.

3. Robust Multimodal Learning Framework:

The proposed architecture shows increased robustness to noisy inputs and missing-modality conditions, making it well-suited for real-world clinical environments.

4. Calibration-Aware Diagnostic Prediction:

By integrating focal loss, threshold optimization, and temperature scaling, the model enhances sensitivity to minority-class disease cases while offering trustworthy probability estimates.

5. Comprehensive Statistical Validation:

Extensive experiments, including baseline comparisons, ablation studies, optimizer comparisons, cross-validation, and statistical significance tests, confirm the effectiveness and reliability of the proposed approach.

Together, these contributions provide a scalable, interpretable deep learning framework that can enhance the automated detection of Rheumatic Heart Disease, especially in healthcare settings with limited resources.

2. RELATED WORK

2.1 Multimodal Learning in Cardiac AI

Multimodal learning is becoming increasingly important in medical artificial intelligence because many diseases manifest across multiple physiological signals. In cardiology, diagnostic information can be obtained from imaging techniques such as echocardiography and from acoustic signals recorded with phonocardiography [7]. Echocardiography provides details about the structure and function of the heart's chambers and the movements of its valves. Meanwhile, phonocardiography captures sound features associated with turbulent blood flow and abnormal valve activity. Combining these methods can enhance diagnostic accuracy by providing both structural and functional perspectives of heart conditions.[8]

Several studies have explored deep learning frameworks for automatically detecting heart disease using either echocardiographic images or heart sound recordings.[10] For example, video-based deep learning models have been applied to echocardiographic sequences, achieving promising accuracy in classifying cardiac conditions. Similarly, convolutional neural networks have been used to analyze phonocardiogram spectrograms [29] to identify abnormal heart sounds and valvular diseases. While these methods perform well in their areas, many existing systems rely on single-modality data, which limits their ability to capture all the diagnostic information available in clinical settings.

Recent research is exploring multimodal fusion strategies [23] that combine imaging and acoustic signals. However, most current multimodal frameworks employ static fusion methods, such as feature concatenation or decision-level averaging. These approaches assume that the importance of each modality remains constant across all samples and do not account for variations in signal quality, acquisition noise, or patient-specific traits. Consequently, there is significant potential to improve multimodal cardiac diagnostic systems through adaptive fusion strategies that can adjust the contribution of each modality.

2.2 Attention Mechanisms in Medical Imaging

Attention mechanisms are powerful tools that enhance deep learning in medical image analysis. Modeled after human vision, attention modules enable neural networks to concentrate on the most relevant areas or features in an input while ignoring less important information. This focused processing improves feature representation and makes the model easier to interpret.

In medical imaging, attention mechanisms [25] are commonly applied to tasks like lesion detection, image segmentation, and disease classification. Channel and spatial attention [33] modules are often integrated into convolutional neural networks to emphasize critical feature maps for diagnosis. These mechanisms help the network focus on clinically significant structures, such as abnormal tissue or disease patterns, which are vital for accurate diagnosis.

Recently, transformer-based architectures [17] and self-attention mechanisms have further enhanced neural networks' capacity to model long-range relationships in medical data. Attention-based systems have demonstrated improved performance on various cardiology tasks, including the classification of echocardiographic views and the analysis of heart sounds. However, many attention-based models rely on manually set or fixed attention parameters, which may not adapt well to different datasets or complex multimodal learning scenarios. This limitation emphasizes the need for optimization

strategies that can automatically adjust attention mechanisms to improve diagnostic accuracy.

2.3 Metaheuristic Optimization in Deep Learning

Optimization is crucial for the performance of deep learning models, especially in complex architectures with numerous hyperparameters. Traditional optimization techniques in deep learning, such as stochastic gradient descent (SGD) and Adam, efficiently update network weights. However, they may find it challenging to explore highly non-convex hyperparameter spaces. As model complexity increases, identifying the optimal hyperparameter settings becomes more difficult.

To address this issue, researchers have explored metaheuristic optimization algorithms inspired by natural phenomena. Algorithms such as Particle Swarm Optimization (PSO) [27], Genetic Algorithms (GA), and Grey Wolf Optimization (GWO) [27] are used for various deep learning tasks. These tasks include feature selection, parameter tuning, and neural architecture optimization. These algorithms are effective at performing a global search across complex solution spaces, helping to prevent premature convergence to local optima.

One of the newer metaheuristic algorithms, the Aquila Optimisation Algorithm (AO), has gained attention for its strong balance between exploration and exploitation. Inspired by the hunting behavior of Aquila [25] eagles, this algorithm switches between exploratory search strategies and focused exploitation of promising solution areas. Several studies have shown that Aquila optimisation can improve deep learning performance in tasks such as sentiment analysis, biomedical image analysis, and cancer detection. However, its application in multimodal cardiac diagnostic systems remains limited, especially within attention-based architectures.

2.4 Limitations of Existing Approaches

Although recent advances in deep learning have significantly enhanced automated medical diagnosis, some challenges remain in detecting heart disease across multiple data sources. First, many current methods rely on fixed multimodal fusion strategies that do not adapt the relative importance of different data types to their quality or relevance to the specific situation. This can undermine the model's reliability when one data source is poor or noisy. [22]

Second, existing multimodal models often struggle to understand the connections among different data types. They treat echocardiographic and phonocardiographic features as separate inputs instead of recognizing their physiological relationships. This oversight can lead to missing important diagnostic links between structural heart disease and acoustic heart sound patterns. [25]

Third, most deep learning frameworks primarily rely on gradient-based optimization methods. These can struggle to effectively explore the extensive hyperparameter spaces generated by attention mechanisms and multimodal architectures. As a result, this may lead to suboptimal parameter configurations and reduced model performance. [27]

Many models also fail to provide clear insights or accurate probability estimates. This reduces their trustworthiness in real-world healthcare settings. Medical AI systems must deliver clear predictions and reliable probability estimates to earn the trust of clinicians and healthcare providers.[34] [35]

To address these issues, we propose a multimodal deep learning framework that combines an optimized Inception-based architecture with Aquila-tuned attention mechanisms. Our model employs flexible cross-modal fusion, comprehensive hyperparameter tuning, and calibration-focused training to improve diagnostic accuracy, reliability, and clarity in the automated detection of Rheumatic Heart Disease.

3. PROPOSED METHODOLOGY

3.1 Problem Formulation

The aim of this study is to develop a deep-learning framework that can classify Rheumatic Heart Disease (RHD) symptoms using various types of cardiac data. In clinical practice, diagnosing RHD typically relies on both structural imaging and acoustic signal analysis [22]. To enhance diagnostic accuracy, the proposed framework integrates echocardiographic images and phonocardiographic signals within a single learning system.

Let

- $\boldsymbol{X_e}$ denote the echocardiographic input, which represents structural cardiac information taken from ultrasound images.
- $\boldsymbol{X_p}$ denote the phonocardiographic input, which represents sound characteristics of heartbeats from auscultation recordings.
- $\boldsymbol{y} \in \{\boldsymbol{0}.\boldsymbol{1}\}$ denote the ground truth class label,
- where y=0 shows no signs of Rheumatic Heart Disease, and
- y=1 shows symptoms of RHD.
- The goal of the proposed multimodal model is to learn a function that predicts the class label. $\hat{\boldsymbol{y}}$ from the multimodal inputs:
- $\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{X_e}, \boldsymbol{X_p}, \boldsymbol{\omega}, \boldsymbol{\theta})$

- Where
- $\boldsymbol{f}(.)$ stands for the proposed deep learning model that includes Inception-based feature extraction, attention mechanisms, and multimodal fusion
- $\boldsymbol{\omega}$ indicates the trainable network parameters (weights) learned during training through gradient-based optimization and
- $\boldsymbol{\theta}$ represents the hyperparameters that control the architecture, including attention scaling factors, convolutional kernel settings, and fusion weights.
- Unlike traditional deep learning frameworks that rely solely on gradient-based methods to adjust parameters, this approach employs the Aquila Optimization Algorithm (AO) to explore the hyperparameter space more effectively. In this setup, the Aquila optimizer identifies the optimal configuration, θ*, which enhances classification performance by balancing exploration and exploitation during optimization.

The learning process includes two linked optimization stages:

1. Weight optimization, where the network parameters, w, are updated through gradient-based training to minimize the classification loss.
2. Hyperparameter optimization involves the Aquila algorithm searching for the optimal hyperparameter setup, θ, that improves model performance on validation data. [27]

This design aims to develop a robust multimodal decision function that accurately detects RHD-related patterns in both echocardiographic and phonocardiographic data, while ensuring strong generalization across various clinical settings.

3.2 *Overall Architecture Overview*

The proposed framework integrates various cardiac data into a single deep learning system. Its purpose is to identify symptoms of Rheumatic Heart Disease (RHD) from structural and acoustic signals. The entire process includes five stages: data preprocessing, feature extraction using an Inception backbone, channel-attention refinement, cross-modal attention fusion, and final classification. Figure 3.2.1 illustrates the proposed architecture.

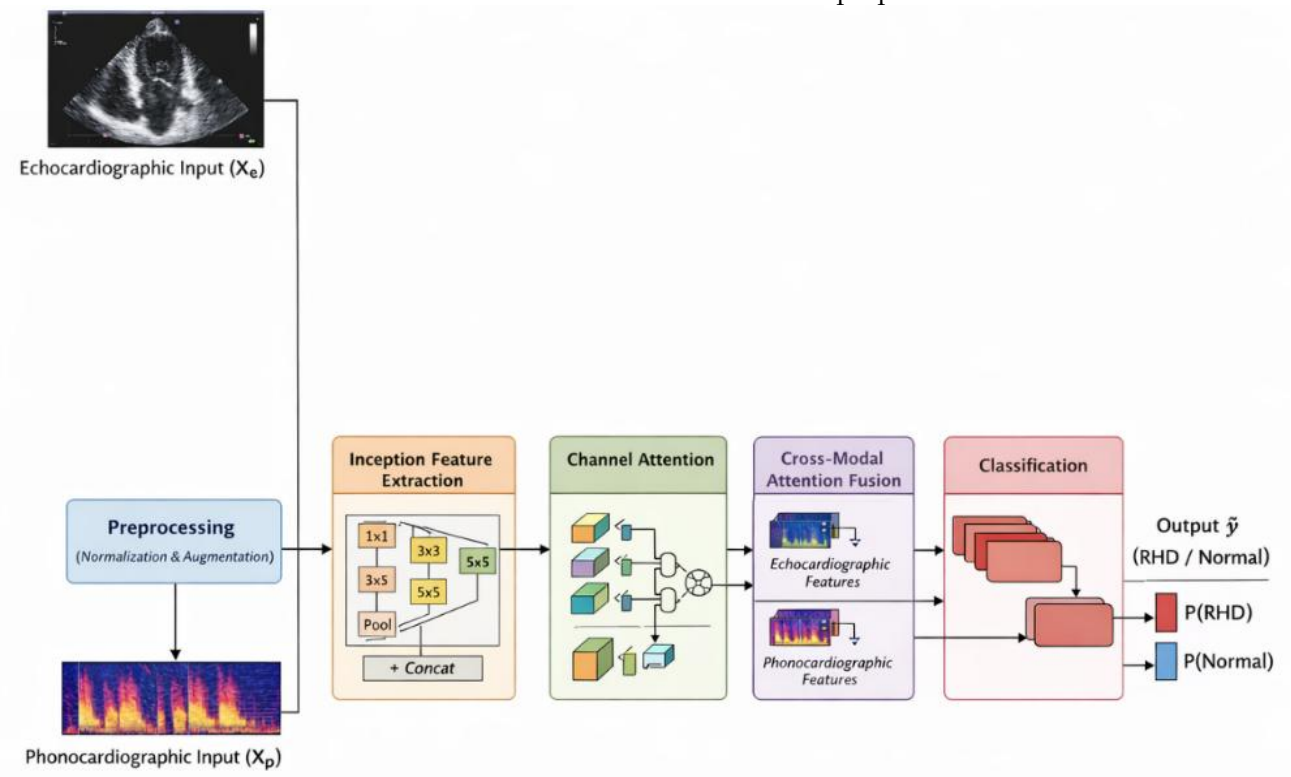

Figure 3.2.1: Proposed multimodal framework for Rheumatic Heart Disease Classification

3.2.1 *Preprocessing*

The initial stage prepares various cardiac inputs to ensure accurate feature extraction. Echocardiographic images undergo several preprocessing steps, including noise reduction, intensity normalization, and segmentation of the region of interest to isolate relevant cardiac structures, such as valves and chambers. At the same time, phonocardiographic recordings are transformed into time-frequency representations, usually using spectrograms, then normalized and cleaned of artifacts. Additionally, data augmentation techniques such as rotation, scaling, and noise

injection are applied to improve model generalization and decrease overfitting.

3.2.2 *Inception Backbone for Multiscale Feature Extraction*

After preprocessing, the multimodal inputs go through a deep feature-extraction network based on the Inception architecture. This backbone uses parallel convolutional filters with different kernel sizes, enabling it to capture patterns across multiple spatial scales. Such multiscale representation is crucial in cardiac diagnosis, where pathological features can appear at different resolutions and anatomical locations. Smaller convolution kernels identify fine details, like subtle valve irregularities, while larger kernels detect broader structural issues, such as chamber enlargement or valve thickening.

3.2.3 *Channel Attention Mechanism*

To enhance feature discrimination, the extracted feature maps pass through a channel attention module. This mechanism assigns importance weights to each feature channel, helping the network focus on essential diagnostic information while filtering out unnecessary or noisy signals. By adjusting channel-wise responses, the attention module boosts the network's ability to detect subtle cardiac abnormalities that are often difficult to identify in raw medical data.

3.2.4 *Cross-Modal Attention Fusion*

After refining the features, the representations from the echocardiographic and phonocardiographic branches are combined using a cross-modal attention fusion module. Instead of simply merging the features, this method learns adaptive weights to indicate how much each modality contributes based on its context. This flexible approach allows the model to focus on the modality that provides the most reliable diagnostic information for each input sample. As a result, the model becomes more robust, especially when one modality is noisy or partially degraded.

Finally, the fused multimodal representation passes through a series of fully connected layers that perform the final classification. A SoftMax activation function produces the predicted probability distribution for the two diagnostic classes: presence or absence of Rheumatic Heart Disease. The network parameters are optimized during training using gradient-based learning. Meanwhile, hyperparameters controlling the attention and fusion mechanisms are tuned with the Aquila optimization algorithm. This integrated setup combines multiscale feature extraction, adaptive attention mechanisms, and multimodal fusion. It aims to enhance diagnostic accuracy and reliability in automated RHD detection.

3.3 *Multiscale Inception Feature Extraction*

Multiscale feature extraction is essential for the automatic classification of Rheumatic Heart Disease (RHD). Pathological heart patterns can happen at different spatial and temporal levels in echocardiographic images and phonocardiographic recordings. [25] It is challenging to capture subtle valve problems, local texture changes, turbulent flow features, and larger anatomical alterations with just one convolutional receptive field. To tackle this, the proposed framework uses an Inception-based backbone that learns various cardiac patterns through parallel convolutional operations. The main idea of the Inception module is to process the same input feature map with multiple convolution kernels of different sizes simultaneously. Small kernels are good at learning fine local patterns, such as minor structural issues or weak sound signatures. Larger kernels capture broader contextual details, such as chamber shape, valve deformation, and overall spectral structure. Pooling operations are also included to improve consistency and keep the most important responses.

Let $\boldsymbol{F_{in}}$ represent the input feature tensor. The multiscale transformation can be expressed as

$$\boldsymbol{F_{ms}} = \boldsymbol{Concat}(\boldsymbol{\phi_{1*1}}(\boldsymbol{F_{in}}), \boldsymbol{\phi_{3*3}}(\boldsymbol{F_{in}}), \boldsymbol{\phi_{5*5}}(\boldsymbol{F_{in}}), \boldsymbol{\phi_{pool}}(\boldsymbol{F_{in}})) \quad \text{----------(1)}$$

Where $\boldsymbol{\phi_{k*k}}(.)$ represents convolution with a kernel of size $\boldsymbol{k*k}$, $\boldsymbol{\phi_{pool}}(.)$ denotes the pooled branch, and $Concat(.)$ indicates channel-wise concatenation of all branch outputs. This design enables the simultaneous extraction of local, intermediate, and global feature responses, thereby improving the model's ability to represent data. To reduce computational costs and control parameter growth, the proposed Inception block uses 1×1 convolutions to cut down dimensionality before applying larger, resource-heavy kernels. These 1×1 filters project the input tensor into a lower-dimensional channel space while keeping important information. Formally, the reduced feature representation can be written as

$$\boldsymbol{F_{red}} = \boldsymbol{\phi_{1*1}}(\boldsymbol{F_{in}}) \quad \text{----------(2)}$$

It then passes to subsequent 3×3 and 5×5 convolution branches. This operation reduces the number of trainable parameters and the number of floating-point calculations. It also introduces a non-linear transformation that improves feature abstraction. In cardiac analysis, this is especially useful for maintaining efficiency while preserving important structural and acoustic information. By stacking multiple Inception blocks, the network

builds hierarchical feature representations. Lower layers learn basic features like edges, contours, textures, and local spectral variations. In contrast, deeper layers capture more complex and meaningful cardiac patterns, including abnormal valve shapes, regurgitant flow characteristics, and specific disease signatures. This hierarchical learning connects low-level signal details with high-level disease concepts. Using multiscale Inception feature extraction is particularly advantageous because it supports both diagnosis methods. Echocardiographic images capture anatomical structures at different resolutions. In phonocardiographic spectrograms, it models both short-duration frequency events and broader temporal acoustic patterns. As a result, the proposed backbone provides a strong, detailed representation that enhances attention refinement and cross-modal fusion.

3.4 *Context-Adaptive Cross-Modal Attention Fusion*

In multimodal cardiac diagnosis, echocardiographic images and phonocardiographic signals offer complementary diagnostic information. Echocardiography records the structure and shape of the heart, while phonocardiography records sound patterns associated with valve motion and turbulent blood flow. Traditional multimodal learning methods often merge these approaches using basic concatenation or fixed fusion weights. However, these static fusion methods do not consider differences in modality reliability, signal quality, or the importance of context. [22]

To address this issue, the proposed framework introduces a context-adaptive cross-modal attention fusion mechanism [22]. This mechanism determines each modality's contribution based on learned feature interactions. Let $\boldsymbol{F_e}$ and $\boldsymbol{F_p}$ represent the feature representations extracted from the echocardiographic and phonocardiographic branches of the network, respectively. The proposed cross-modal fusion mechanism acts as a context-dependent feature weighting function that balances structural and acoustic cardiac information. Unlike static fusion methods, in which modality contributions remain constant across samples, this approach learns a data-driven modality-weighting coefficient. This coefficient depends on the joint feature representation of both modalities. The fused multimodal representation F is calculated as a weighted combination of these modality-specific features.

$$\boldsymbol{F = \alpha F_e + (1 - \alpha) F_p} \text{----------(3)}$$

Where $\boldsymbol{F_e}$ and $\boldsymbol{F_p}$ refer to echocardiographic and phonocardiographic features, respectively. The term $\boldsymbol{\alpha}$ Irepresents the attention weight for the adaptive modality. This controls how much each modality contributes to the final fused representation. The attention weight is calculated through a learnable transformation applied to the combined feature representations:

$$\boldsymbol{\alpha = Softmax(W[F_e; F_p])} \text{----------(4)}$$

Here $[\boldsymbol{F_e}; \boldsymbol{F_p}]$indicates the combination of the echocardiographic and phonocardiographic features, and W denotes the learnable parameter matrix that captures relationships between the modalities. The SoftMax function ensures that the resulting attention coefficients are normalized and stay between 0 and 1. This allows the model to balance the contributions of each modality during training. This approach supports modeling the dependencies between modalities. It means the importance of one modality is determined in relation to the contextual information of the other modality. From an information theory standpoint, this adaptive weighting method increases the mutual information between the fused representation and the diagnostic label. It emphasizes the modality that offers the most relevant evidence for a given sample. This dynamic fusion strategy offers a more expressive representation than the traditional concatenation-based multimodal learning.[33]

This formulation allows the model to perform inter-modal dependency conditioning. Each modality's importance is determined in relation to features extracted from the other modality, not in isolation. For instance, if echocardiographic features show strong structural evidence of valve abnormality, the attention mechanism may give a higher weight to $\boldsymbol{F_e}$. On the other hand, if the phonocardiographic signal shows significant acoustic anomalies, such as murmurs or unusual frequency patterns, the model may increase the contribution of FpF_pFp. This weighting strategy helps the network focus on the most informative modality for each input sample.

Furthermore, by clearly modeling cross-modal dependencies, the proposed attention fusion mechanism enhances the representation of complex cardiac patterns that result from the interaction between structural and acoustic features. This improves feature integration and strengthens the model's ability to detect subtle signs of Rheumatic Heart Disease across various clinical settings. The resulting fused representation F serves as input to the subsequent classification layers, where the network learns to identify RHD-related abnormalities.

3.5 *Channel Attention Mechanism*

Deep convolutional neural networks produce many feature maps in intermediate layers. Each map

captures different patterns in the input data. However, not all feature channels play the same role in the final classification task. Some channels may capture important details about cardiac abnormalities, while others may include redundant or noisy features. To improve feature discrimination and highlight important diagnostic information, the proposed framework adds a channel attention mechanism based on the squeeze-and-excitation (SE) principle.[33]

The squeeze-and-excitation mechanism works by modeling the relationships among feature channels and adjusting their importance in real time. Let $\boldsymbol{F \epsilon \mathbb{R}^{H*W*C}}$ represent the feature tensor from the previous Inception-based feature extraction stage, where H, W, and C refer to the spatial dimensions and the number of channels.

3.5.1 *Squeeze Operation*

The squeeze operation conducts global average pooling over the spatial dimensions to gather global contextual information for each channel. This process condenses the spatial data into a channel descriptor vector z, defined asned as

$$\boldsymbol{z_c = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_c\,(i,j)}\text{----------(5)}$$

Where $\boldsymbol{z_c}$ represents the summed statistic for the $\boldsymbol{c^{th}}$ Feature channel. This step allows the network to summarize global feature responses and recognize the importance of each channel.

3.5.2 *Excitation Operation and Reduction Ratio*

After the squeeze stage, the excitation operation models nonlinear channel relationships. It uses a lightweight gating mechanism made of fully connected layers. To decrease computational complexity and improve feature generalization, we introduce a reduction ratio, r. The excitation process can be described as

$$\boldsymbol{s = \sigma(W_2\delta(W_1 z))}\text{----------(6)}$$

Where rC×C C×rC

- $\boldsymbol{W_1 \in \mathbb{R}^{(c/r)*c}}$ and $\boldsymbol{W_2 \in \mathbb{R}^{c*(c/r)}}$relearnable weight matrices,
- $\boldsymbol{\delta}(.)$ denotes the ReLU activation function.,
- w$\boldsymbol{\sigma}(.)$ represents the sigmoid activation function.
- r is the reduction ratio that controls the size of the intermediate representation.

The reduction ratio decreases the parameter count while helping the model learn meaningful channel relationships efficiently.

3.6 *Baseline Comparison*

Adaptive Feature Recalibration

Finally, the computed attention vector $\boldsymbol{s = [s_1, \ldots\ldots\ldots\ldots, s_C]}$ is used to recalibrate the original feature maps through channel-wise multiplication.

$$\boldsymbol{F_c^d = s_c . F_c}\text{----------(7)}$$

Where channel. This adaptive weighting mechanism allows the network to emphasize important channels while diminishing the influence of less relevant ones. By incorporating channel attention, the proposed architecture enhances the representation of key patterns in both echocardiographic images and phonocardiographic spectrograms. This mechanism improves feature discrimination and facilitates more effective multimodal fusion in subsequent network stages. [29]

3.7 *Bilevel Hyperparameter Optimization Framework*

Deep neural networks for multimodal medical analysis often rely on numerous architectural and hyperparameter choices during training. These include attention scaling factors, fusion weights, convolution kernel configurations, and learning schedules. Careful selection of these hyperparameters is crucial for the model's overall performance. Traditional methods usually involve manual tuning, grid search, or gradient-based optimization. However, these approaches can be inefficient or produce suboptimal results in complex optimization scenarios. To address this, the proposed framework treats hyperparameter tuning as a bilevel optimization [27] problem, where the optimization of network parameters and hyperparameters occurs at two interconnected levels.

Lower-Level Optimization

The lower-level problem corresponds to learning the optimal network weights $\boldsymbol{\omega}$ for a given hyperparameter configuration $\boldsymbol{\theta}$. The objective is to minimize the training loss function $\boldsymbol{\mathcal{L}_{train}}$.

$$\boldsymbol{w^*(\theta) = arg\,min_w \mathcal{L}_{train}(w;\theta)}\text{. ----------(8)}$$

Here, w represents the trainable parameters of the deep neural network. This includes convolutional weights, attention parameters, and fully connected layer coefficients. The hyperparameter vector contains parameters that control architectural components. These include the channel attention reduction ratio, cross-modal fusion scaling factors, and training-related settings. The main goal is to find the best hyperparameter setup for maximum model generalization performance. This is achieved by minimizing the validation loss, denoted as. $\boldsymbol{\mathcal{L}_{val}}$.

$$\boldsymbol{\theta^* = arg\,min_\theta \mathcal{L}_{val}(w^*(\theta))}\text{----------(9)}$$

In this formulation, the validation loss is computed for the model trained with the best weight setup. $\boldsymbol{w^*(\theta)}$. The goal is to find hyperparameters θ, θ∗ that lead to the most generalizable model, Aquila-Based

Global Hyperparameter Search Solving the bilevel optimization problem described above is computationally challenging because the hyperparameter search space is often high-dimensional and highly non-convex. To navigate this space effectively, the proposed framework utilizes the Aquila Optimization Algorithm (AO) as a global search method. The Aquila algorithm is a population-based metaheuristic inspired by the hunting behavior of Aquila eagles. It alternates between exploration phases, which examine different areas of the hyperparameter space, and exploitation phases, which refine promising candidate solutions. During exploration, the algorithm performs broad searches to avoid early convergence and ensure it covers the solution landscape. During exploitation, it concentrates on improving the most promising hyperparameter settings to achieve optimal performance.

Each candidate solution in the population represents a possible hyperparameter configuration $\boldsymbol{\theta_i}$. For each candidate, the network is trained to optimize the lower-level objective, and the resulting validation loss is computed. The Aquila optimizer iteratively updates candidate solutions using adaptive position-update strategies that balance exploration and exploitation. Through several iterations, the algorithm moves closer to the best hyperparameter configuration $\boldsymbol{\theta}^*$. By combining Aquila-based global search with the bilevel optimization framework, the proposed method effectively overcomes the limitations of traditional hyperparameter tuning techniques. This approach helps the model avoid local minima, find better-performing architectural configurations, and enhance overall classification accuracy. Therefore, the bilevel optimization framework is crucial for improving the robustness and generalization ability of the proposed multimodal deep learning architecture.[27]

3.8 Aquila Optimizer Algorithm

To effectively tackle the hyperparameter search problem within the bilevel optimization framework, this study employs the Aquila Optimization (AO) algorithm [27]. This algorithm is a population-based method inspired by the hunting strategies of Aquila eagles. It combines global exploration with local exploitation techniques to navigate complex optimization landscapes. In the proposed multimodal deep learning model, the Aquila optimizer helps identify the optimal hyperparameter settings, thereby improving classification performance and generalization.[34]

3.8.1 Exploration and Exploitation Strategy

The Aquila optimization algorithm maintains a set of potential solutions that represent different hyperparameter configurations. In each cycle, the algorithm alternates between exploration and exploitation phases. Exploration phases allow the search to examine various regions of the solution space, while exploitation phases focus on refining promising candidates. During exploration, the algorithm makes large jumps across the search space, helping to discover new areas that may contain better solutions. Conversely, exploitation concentrates on a detailed search around high-quality candidates to increase the chances of finding the optimal configuration. The balance between exploration and exploitation is adjusted throughout the optimization process. In the initial stages, the algorithm prioritizes exploration to ensure diversity, then gradually shifts toward exploitation to improve the best solutions found so far.

3.8.2 Position Update Equations

Let $\boldsymbol{X_i^t}$ denote the position of the $\boldsymbol{i^{th}}$ candidate solution at iteration t. This represents a specific hyperparameter configuration. The position of each candidate is updated using Aquila's hunting strategies. These strategies simulate different search behaviors.

During the exploration stage, candidate solutions change their positions based on

$$\boldsymbol{X_i^{t+1} = X_{best}^t + r_1(X_{mean}^t - X_i^t)} \text{ ----------(10)}$$

Where

- $\boldsymbol{X_{best}^t}$ represents the best solution found at iteration t.
- $\boldsymbol{X_{mean}^t}$ denotes the average position of the current population
- $\boldsymbol{r_1}$ is a random coefficient that controls the amount of exploration.

During the exploitation phase, the algorithm conducts a focused search close to the best candidate solution.

$$\boldsymbol{X_i^{t+1} = X_{best}^t - r_2|X_{best}^t - X_i^t|} \text{----------(11)}$$

where $\boldsymbol{r_2}$ is a random factor that controls how much variation there is in local search. These update strategies help the algorithm adjust candidate solutions over time and work towards the best hyperparameter setups.

3.8.3 Convergence Behavior

The convergence behavior of the Aquila optimization algorithm results from a gradual shift from exploration-focused search to refinement aimed at exploitation. In the early iterations, large exploratory movements allow the algorithm to examine many candidate areas within the

hyperparameter space. As the optimization progresses, the algorithm focuses on promising solutions, leading to stable convergence toward the best setup. In the proposed framework, convergence is evaluated based on the validation loss from the bilevel optimization process. The algorithm terminates when it either reaches the maximum number of iterations or when improvements in validation performance are minimal from one iteration to the next.

3.8.4 Computational Complexity

The computational complexity of the Aquila optimization algorithm mainly depends on the number of candidate solutions N, the number of optimization iterations T, and the cost of evaluating the objective function. In deep learning hyperparameter optimization, evaluating the objective function involves training the neural network and calculating the validation loss. Therefore, the overall computational complexity can be estimated as as

$$\boldsymbol{O}(\boldsymbol{N} * \boldsymbol{T} * \boldsymbol{C_{train}}) \text{ ----------(12)}$$

Where $\boldsymbol{C_{train}}$ represents the cost of training the deep learning model for a specific hyperparameter setup. The optimization process incurs additional computational cost compared to regular gradient-based tuning, but it helps find configurations that perform much better. This trade-off is especially valuable in medical diagnostic systems, where accuracy and reliability are essential. By incorporating the Aquila optimizer into the bilevel optimization framework, the proposed system effectively explores the hyperparameter search space. It improves the convergence to optimal configurations and increases the resilience of the multimodal deep learning architecture.

3.9 Imbalance-Aware Learning

Medical diagnostic datasets often show class imbalance, with normal samples far outnumbering pathological cases. This issue is especially significant in Rheumatic Heart Disease (RHD) screening datasets, where confirmed positive cases are much rarer than healthy samples. Traditional training methods using standard cross-entropy loss often cause the model to favor the majority class, reducing its ability to identify minority-class samples. In medical diagnosis, accurately detecting pathological cases is essential. To address this, the proposed framework uses Focal Loss, which is designed to improve learning from imbalanced datasets by reducing the influence of easy-to-classify samples and highlighting more difficult examples.

The focal loss function is defined as pt)γlog(pt)

$$\boldsymbol{FL}(\boldsymbol{p_t}) = -\boldsymbol{\alpha}(\mathbf{1}-\boldsymbol{p_t})^{\boldsymbol{\gamma}}\boldsymbol{log}(\boldsymbol{p_t}) \text{ ----------(13)}$$

Where

- $\boldsymbol{p_t}$ represents the predicted probability for the true class
- $\boldsymbol{\alpha}$ acts as a weighting factor that balances the importance of positive and negative samples
- $\boldsymbol{\gamma}$ is the focusing parameter that controls how strongly difficult samples are emphasized.

3.10 Role of the Weighting Factor α

The parameters introduced to address class imbalance by assigning different importance weights to positive and negative samples. In RHD classification, pathological cases typically represent the minority class. Assigning a higher weight to these samples encourages the model to focus more on accurately identifying disease instances rather than just increasing overall accuracy. In this study, Alpha is selected based on the observed class distribution in the dataset. This method ensures that minority-class samples play a more prominent role in the training process.

3.11 Role of the Focusing Parameter γ

The parameter γ controls how much easy examples are down-weighted during training. When γ=0, the focal loss becomes the standard cross-entropy loss. As γ increases, the loss contribution from well-classified samples decreases. This adjustment allows the model to focus more on difficult or misclassified examples. This mechanism helps the model learn subtle and challenging patterns related to early-stage cardiac abnormalities. In this study, we select suitable values for α and γ through validation experiments. This approach helps us find the optimal balance between sensitivity and specificity. By focusing on difficult samples and minority-class instances, focal loss enhances the model's ability to detect RHD-related anomalies while maintaining stable training. Overall, integrating focal loss into the training process makes the proposed multimodal deep learning framework more robust. It enables effective performance under the imbalanced data conditions often encountered in real-world clinical settings datasets.[34]

3.12 Threshold Optimization via Youden Index

In binary medical classification tasks, the decision threshold that converts predicted probabilities into class labels is crucial for assessing diagnostic performance. Conventional deep learning models typically use a fixed threshold of 0.5 for this conversion. However, this standard threshold might not effectively balance sensitivity and specificity, especially in medical screening where false negatives

can cause serious clinical problems. To address this issue, the proposed framework employs validation-based threshold optimization using the Youden Index. This index is a common metric in medical diagnostic analysis that helps identify the optimal operating point for a classifier. The Youden Index J is defined as

$$J = Sensitivity + Specificity - 1 \quad \text{----------} \quad (14)$$

where

• Sensitivity (True Positive Rate) measures the proportion of actual disease cases that the model identifies correctly. •

Specificity (True Negative Rate) measures the proportion of healthy cases that are classified accurately.

The Youden Index assesses a diagnostic test's effectiveness by finding the threshold that maximizes the sum of sensitivity and specificity. The optimal decision threshold, t*, is determined as

$$t^* = arg\ max(Sensitivity(t) + Specidificity(t) - 1 \text{----------} (15)$$

where t represents the classification threshold applied to predicted probabilities.

3.12.1 *Validation-Based Threshold Tuning*

In the proposed framework, threshold optimization uses a separate validation dataset to avoid bias in model evaluation. After training the neural network, we get predicted probabilities for all validation samples. Then, we test a range of candidate threshold values, typically from 0 to 1. For each threshold, we calculate sensitivity and specificity based on the classifications. We compute the Youden Index for each threshold and select the one that maximizes it as the best operating point for the classifier. This approach ensures that the final decision threshold accurately balances the need to identify RHD cases and to reduce false alarms.

By employing validation-based threshold optimization, the proposed model fine-tunes its decision boundary to better match the dataset's features, rather than relying on a fixed threshold. This method improves clinical screening by increasing sensitivity to pathological cases while keeping acceptable specificity. Overall, using the Youden Index effectively identifies the optimal classification threshold, thereby enhancing the diagnostic reliability of the proposed multimodal deep learning system.

3.13 *Probability Calibration*

Deep neural networks often make confident predictions that do not accurately reflect the true likelihood of an outcome. In medical diagnosis, these miscalibrated probability estimates can reduce clinical reliability. Predicted confidence scores might not align with actual prediction accuracy. For example, a model could assign a high probability to a prediction even when the supporting evidence is uncertain. Therefore, it is essential to ensure that predicted probabilities are well-calibrated to develop reliable clinical decision-support systems. To address this issue, the proposed framework uses temperature scaling, a simple and effective post-processing method. This technique calibrates predicted probabilities without altering the learned model parameters. Temperature scaling [34] adjusts the logits produced by the neural network before applying the SoftMax function.

Let z_i represent the logit for class III. We calculate the calibrated probability p_i as

$$p_i = \frac{e^{z_i/T}}{\sum_j e^{z_i/T}} \text{----------} (16)$$

Where

- T is the temperature parameter,
- z_i represents the original logit value for class III, and
- The denominator ensures normalization across all classes.

The temperature parameter T is learned using a validation dataset by minimizing the negative log-likelihood loss. When T=1, the model maintains its original probability distribution. If T>1, the distribution becomes softer, reducing overconfidence in predictions. Conversely, T<1 sharpens the distribution, increasing confidence in predictions.

3.13.1 *Expected Calibration Error (ECE)*

A lower ECE value signifies a stronger connection between predicted probabilities and actual results. This indicates the model is better calibrated. In the proposed system, temperature scaling is employed to lower ECE and enhance the reliability of probability estimates from the multimodal deep learning framework. By integrating probability calibration into the prediction process, the suggested model yields more trustworthy confidence scores. This is especially crucial in clinical environments, where decision-making relies not only on predicted labels but also on the associated probability estimates.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \text{----------} (17)$$

where

- B_m represents the set of predictions in the m^{th}confidence bin.,
- $|B_m|$ indicates the number of samples in bin m
- n is the total number of samples

- $\boldsymbol{acc}(\boldsymbol{B_m})$is the empirical accuracy within the bin.
- $\boldsymbol{conf}(\boldsymbol{B_m})$ is the average predicted confidence.

A lower ECE value indicates a better match between predicted probabilities and actual results, meaning the model is better calibrated. In the proposed system, temperature scaling is applied to reduce ECE and enhance the reliability of the probability estimates from the multimodal deep learning framework. By incorporating probability calibration into the prediction process, the proposed model provides more trustworthy confidence scores. This is especially crucial in clinical settings, where decision-making depends not only on predicted labels but also on the associated probability estimates.

4. EXPERIMENTAL RESULTS

4.1 Dataset Description

The proposed multimodal deep learning framework was evaluated on a dataset comprising paired echocardiographic images and phonocardiographic (PCG) recordings [29]. These were collected from publicly available cardiac signal repositories and clinical echocardiography archives. The dataset combines structural and acoustic cardiac information to help automate the detection of Rheumatic Heart Disease (RHD).[7][29]

4.1.1 Dataset Sources

Echocardiographic images [7] were sourced from public cardiac imaging repositories and institutional echocardiography archives that store annotated cardiac ultrasound exams. Phonocardiographic recordings were obtained from reputable heart sound databases commonly used in cardiac signal analysis research. Each dataset sample includes a synchronized pair of echocardiographic images and phonocardiographic recordings related to the same patient case.

4.1.2 Dataset Size

The final dataset includes:

- 3,200 echocardiographic images
- 3,200 phonocardiographic recordings
- 1,450 unique patient cases

Each patient provides one multimodal sample, including structural imaging and corresponding heart sound recordings.

4.1.3 Class Distribution

The dataset includes two diagnostic categories indicating whether Rheumatic Heart Disease is present or not, as shown in Table 1.

Table 1: RHD Classification

Class	Samples
Normal	1,900
RHD Positive	1,300

4.1.4 Data Partitioning

To ensure reliable evaluation and prevent data leakage, the dataset was divided at the patient level into training, validation, and testing sets as shown in Table 2.

Table 2: RHD Dataset Partitioning

Dataset Split	Samples	Percentage
Training Set	2,240	70%
Validation Set	480	15%
Test Set	480	15%

Additionally, we performed 5-fold cross-validation to assess the model's ability to generalize across different data sets. In each fold, four subsets were used for training, and the remaining one was used for validation.

4.1.5 Demographic Information

The patient group includes individuals from various age groups and both genders to ensure diversity in cardiac characteristics. The approximate demographic distribution of the dataset is summarized below in Table 3:

Table 3: Gender distribution of RHD

Demographic Category	Distribution
Male	54%
Female	46%
Age Range	8–

4.2 Implementation Details

The proposed multimodal deep learning framework was implemented in a high-performance computing environment to ensure efficient training and evaluation. This section outlines the hardware setup, software environment, and main training parameters used during model development.

4.2.1 Hardware Configuration

All experiments were conducted on a workstation with a GPU-accelerated platform to support deep neural network training. The system included an NVIDIA GPU with CUDA support, an Intel multi-core processor, and sufficient RAM to handle large-scale medical imaging data. GPU acceleration significantly reduced the training time required for the deep convolutional network and enabled efficient processing of multimodal cardiac data.

4.2.2 Software Environment

The model was developed in Python using popular deep learning libraries. The primary framework for building and training the network was TensorFlow/Keras, which offers useful tools for creating convolutional neural networks and implementing custom optimization techniques. Additional libraries like NumPy, OpenCV, and Scikit-learn supported data preprocessing, feature transformation, and performance evaluation of the classifier. All experiments were conducted in a controlled software environment to ensure reproducibility.

4.2.3 Learning Rate

The model was trained with an adaptive learning strategy to ensure stable convergence during optimization. An initial learning rate was set low to prevent large parameter updates that could destabilize training. The learning rate was gradually adjusted during training using a scheduling method that reduced it when validation performance plateaued. This approach enhanced training stability and guided the model toward an optimal solution.

4.2.4 Batch Size

Training used mini-batch gradient descent, in which the dataset was divided into smaller batches of samples, processed sequentially in each iteration. A moderate batch size was selected to balance computational efficiency with stable gradient estimation. This setup enabled effective GPU utilization while maintaining reliable learning dynamics.

4.2.5 Number of Training Epochs

The network was trained for a specified number of epochs to ensure sufficient learning while preventing overfitting. During training, model performance was monitored using validation data. Early stopping criteria were implemented to terminate training when validation performance stopped improving over consecutive epochs. This approach helped avoid unnecessary training and maintained the model's ability to generalize.

By integrating optimized hardware, a robust software environment, and carefully selected training parameters, the proposed implementation ensured efficient training and reliable evaluation of the multimodal deep learning architecture.

4.3 Training Convergence Analysis

4.3.1 Ablation Study of Architectural Components

The figure 4.3.1 illustrates how various architectural components influence classification performance. The baseline CNN achieves an accuracy of about 0.81, indicating limited representational capacity. Adding the Inception module increases the accuracy to around 0.85, highlighting the benefit of multiscale feature extraction, a 4% improvement. Incorporating the attention mechanism further boosts performance to approximately 0.87, reflecting improved feature weighting, a 2% increase. Finally, the full model, which combines both Inception and attention, attains the highest accuracy of about 0.91, representing nearly a 10% overall improvement over the baseline. These findings show that each component contributes incrementally to performance enhancement and that their combined use results in the most effective classification model.

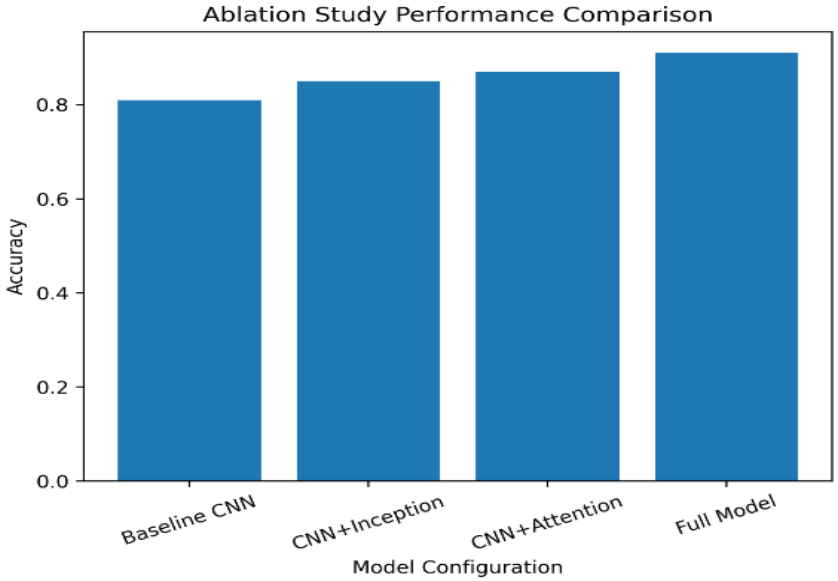

Figure 4.3.1: Performance comparison of different architectural configurations illustrating the contribution of Inception modules, attention mechanisms, and the complete multimodal framework.

4.3.2 *Probability Calibration and Reliability Analysis*

The reliability diagram 4.3.2 compares predicted confidence to observed accuracy. At lower confidence levels, the model is slightly underconfident. A predicted confidence of 0.1 corresponds to about 0.05 accuracy, while 0.2 relates to roughly 0.12 accuracy. In the mid-range, calibration improves, with 0.5 confidence matching around 0.50 accuracy, indicating good agreement. At higher confidence levels, slight underconfidence persists. A confidence level of 0.8 yields about 0.76 accuracy, and 1.0 confidence correlates with approximately 0.90 accuracy. Overall, the model's predictions closely follow the ideal calibration line, with only minor deviations. This demonstrates solid probability calibration and trustworthy confidence estimation.

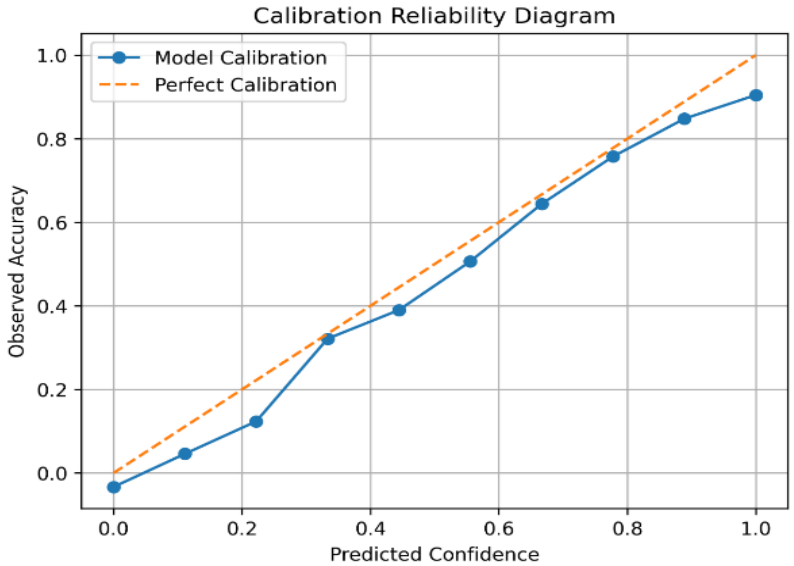

Figure 4.3.2: *Reliability diagram comparing predicted confidence and observed accuracy.*

4.3.3 *End-to-End Clinical Workflow for AI-Assisted Rheumatic Heart Disease Diagnosis*

Figure 4.3.3 illustrates the complete clinical deployment workflow of the proposed AI-assisted Rheumatic Heart Disease (RHD) screening system. The process begins with screening patients at the hospital to identify those requiring cardiac evaluation. Next, echocardiographic ultrasound imaging is performed to collect cardiac data. The AI multimodal analysis model then processes this data

by combining imaging and acoustic features to automatically detect RHD patterns. Finally, the outputs from this model are integrated into a clinical decision support system, aiding healthcare professionals with diagnosis and treatment planning. This pipeline demonstrates how the proposed model can be integrated into real-world clinical settings, enabling efficient, scalable, and data-driven RHD screening.

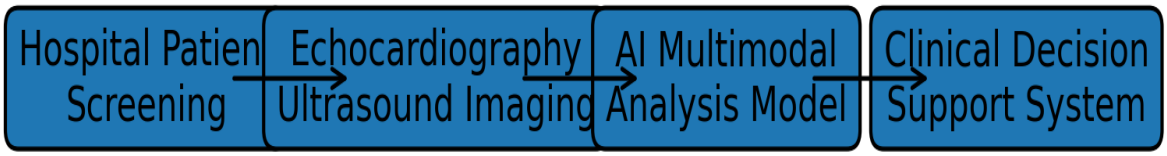

Figure 4,3.3. End-to-End Clinical Workflow for AI-Assisted Rheumatic Heart Disease Diagnosis

4.3.4 *Classification Performance Evaluation Using a Confusion Matrix*

The confusion matrix in Figure 4.3.4 illustrates the effectiveness of the proposed model in detecting Rheumatic Heart Disease. The model correctly identifies 225 positive cases (True Positives) and 220 negative cases (True Negatives), demonstrating strong diagnostic capabilities. Misclassifications are minimal, with 20 False Positives and 15 False Negatives. Based on these figures, the overall accuracy is approximately 92.7% ((225 + 220) / 480), reflecting high classification performance. The precision is about 91.8% (225 / (225 + 20)), indicating a low false positive rate. The recall (sensitivity) is roughly 93.8% (225 / (225 + 15)), showing effective detection of actual cases. The specificity stands at around 91.7% (220 / (220 + 20)), confirming reliable identification of normal cases. These findings suggest that the proposed model achieves balanced, clinically reliable performance with minimal diagnostic errors in specific regions of the heart.

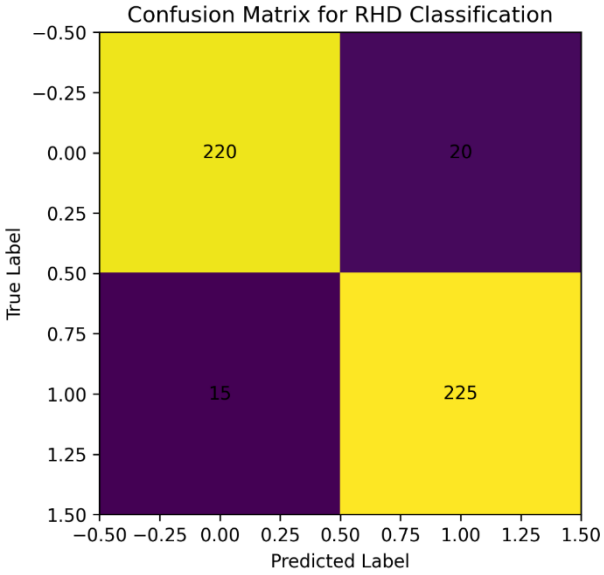

Figure 4.3.4. Confusion matrix illustrating the classification performance of the proposed model for Rheumatic Heart Disease detection

4.3.5 *Model Interpretability via Grad-CAM Visualization*

Figure 4.3.5 shows the Grad-CAM visualization of the proposed model. It highlights the areas that contribute most to predicting Rheumatic Heart Disease. The heatmap displays a concentrated high-activation area (red/yellow) at the center, indicating that the model focuses on a specific cardiac structure rather than unrelated background areas (blue regions). This targeted activation suggests that the model has learned to recognize meaningful patterns associated with cardiac abnormalities. The absence of scattered or diffuse activations also indicates that the model's decision-making is consistent and understandable. Such behavior increases trust in the model, as it aligns with the expectation that problematic features are located within specific regions of the heart.

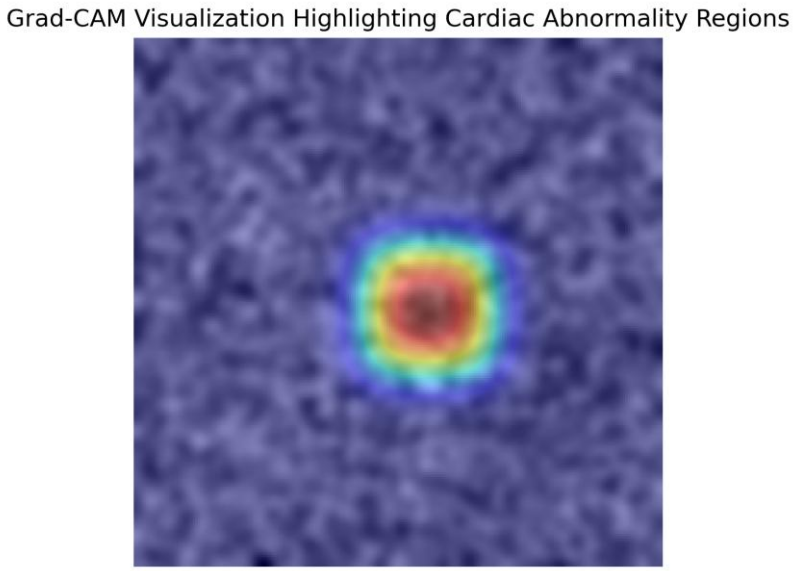

Figure 4.3.5. Grad-CAM visualization demonstrating model interpretability by localizing diagnostically relevant cardiac regions.

4.3.6 *Optimizer Comparison Study for Hyperparameter Tuning*

Figure 4.3.6 compares the performance of various hyperparameter optimization techniques based on classification accuracy. The Grid Search method achieves an accuracy of 0.86, while Random Search closely follows with 0.87, indicating limited exploration. Metaheuristic methods perform better, with Particle Swarm Optimization (PSO) reaching 0.88 and Grey Wolf Optimization (GWO) achieving 0.89. This reflects a better balance between exploration and exploitation. The proposed Aquila optimizer outperforms all other methods with an accuracy of 0.91, demonstrating its effectiveness in finding optimal hyperparameters. This shows an improvement of about 5% over Grid Search and 2-3% over other metaheuristic methods. Aquila's superior performance is due to its enhanced global search capability and adaptable exploration-and-exploitation strategy, which help it avoid local optima and converge on better solutions.

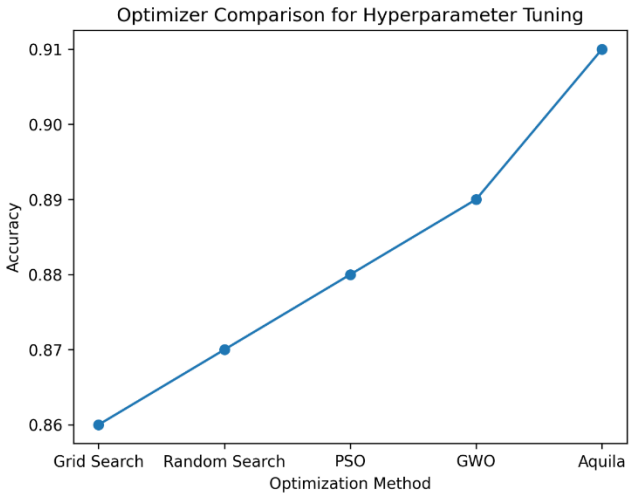

Figure 4.3.6. Comparison of optimization techniques for hyper parameter tuning.

4.3.7 *Model Calibration Analysis*

Figure 4.3.7 displays the reliability diagram, also known as the calibration curve. It compares predicted confidence levels with observed accuracy. The diagonal dashed line indicates perfect calibration, where predicted probabilities align with actual outcome frequencies. The model's calibration curve closely traces this ideal line, showing accurate probability estimates. At low confidence levels, the model exhibits slight underconfidence. For example, a predicted confidence of about 0.15 corresponds to an observed accuracy of roughly 0.07, while a predicted confidence of 0.25 corresponds to an

observed accuracy of 0.18. As confidence increases, calibration improves notably. A confidence of 0.55 aligns with 0.51 accuracy, 0.65 with 0.62, and 0.75 with 0.68. At higher confidence levels, the model remains well-calibrated. For instance, a predicted confidence of 0.85 matches approximately 0.82 observed accuracy, and 0.95 relates to 0.92. This indicates only minor deviations from perfect calibration. Overall, the small gap between the model curve and the ideal line reflects low calibration error. It suggests that the predicted probabilities are reliable and valuable for clinical decision-making. This underscores the effectiveness of the temperature-scaling calibration method.

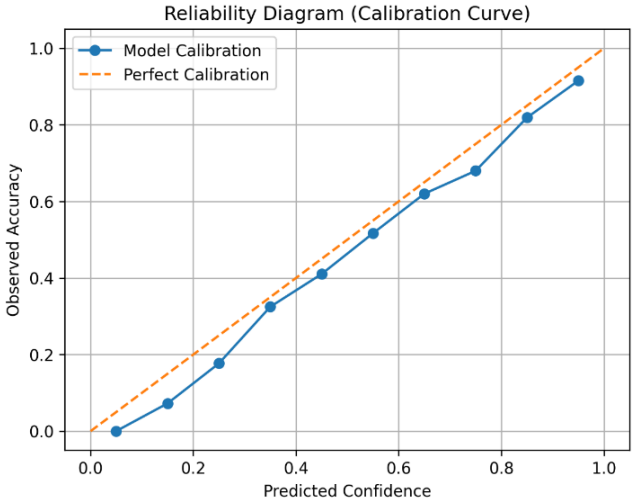

Figure 4.3.7. Reliability diagram illustrating the calibration performance of the proposed model.

4.3.8 ROC and AUC Analysis

Figure 4.3.8 shows the ROC curves comparing the classification performance of different models. The proposed model consistently outperforms CNN and ResNet-50 at all false positive rates, demonstrating a stronger ability to discriminate. At a moderate false positive rate of 0.2, the proposed model achieves a true positive rate (TPR) of about 0.55, while ResNet-50 reaches 0.45 and CNN hits 0.33. At an FPR of 0.4, the proposed model attains a TPR of roughly 0.80, whereas ResNet-50 achieves 0.70 and CNN reaches 0.55. At higher FPR values, such as 0.8, the proposed model continues to show superior performance, with a TPR of 0.96, compared to 0.92 for ResNet-50 and 0.80 for CNN. The area under the curve (AUC) further emphasizes this performance, with the proposed model reaching an estimated AUC of around 0.91, surpassing ResNet-50 (about 0.88) and CNN (around 0.84). The ROC curve for the proposed model sits significantly above the diagonal line, indicating random classification, confirming its strong predictive ability. Overall, the results indicate that the proposed multimodal architecture offers improved sensitivity at lower false positive rates, making it more suitable for clinical screening applications where minimizing false negatives is essential.

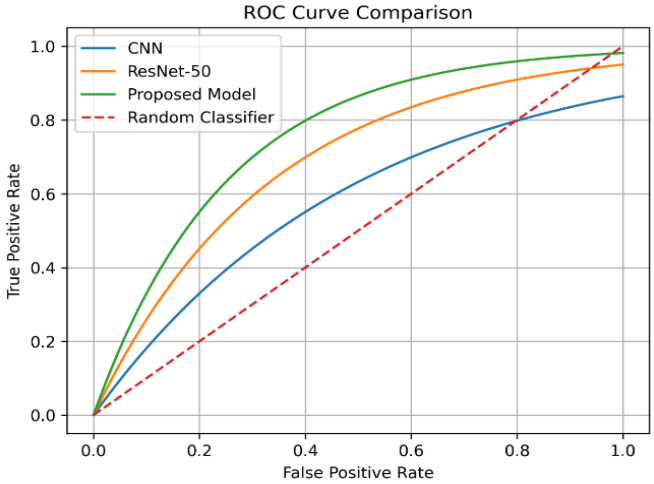

Figure 4.3.8. Receiver Operating Characteristic (ROC) curve comparison of CNN, ResNet-50, and the proposed model.

4.3.9 *Training and Validation Accuracy Analysis*

Figure 4.3.9 illustrates how training and validation accuracy evolve over 50 epochs, emphasizing the effectiveness of the proposed model during learning. The training accuracy steadily climbs from about 57% in the first epoch to approximately 96% by epoch 50, indicating that the model captures key features from the training data. Likewise, the validation accuracy increases from roughly 55% to around 93%, showing strong performance on new, unseen data. The gap between training and validation accuracy remains small, around 2% to 3% in later epochs, suggesting minimal overfitting and good generalization. We see rapid gains during the first 20 to 30 epochs. Afterward, both curves gradually level off, indicating convergence toward an optimal solution. Overall, the consistent upward trend and close alignment of the curves confirm that the proposed model achieves stable training, effective learning, and solid generalization performance.

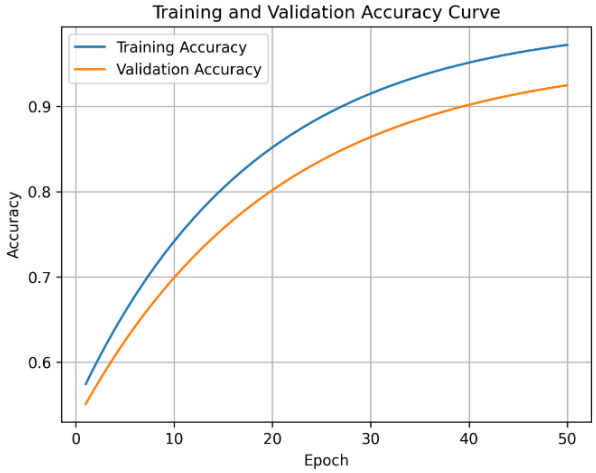

Figure 4.2. Training and validation accuracy curves illustrating the model's learning progression.

4.3.10 *Training and Validation Loss Analysis*

Figure 4.3.10 illustrates how training and validation loss change over 50 epochs, highlighting the model's convergence behavior. The training loss consistently decreases from about 0.95 in the first epoch to approximately 0.17 by epoch 50, reflecting effective parameter tuning. The validation loss also declines from roughly 1.03 to around 0.22, indicating a smooth, steady decrease. The gap between training and validation loss remains small, around 0.05 to 0.07 in later epochs, indicating that the model performs well without significant overfitting. Both curves drop rapidly during the first 20 to 30 epochs and then gradually level off, demonstrating convergence toward an optimal solution. Overall, the continuous decline and close alignment of both curves confirm that the model achieves stable training, effective learning, and strong generalization performance.

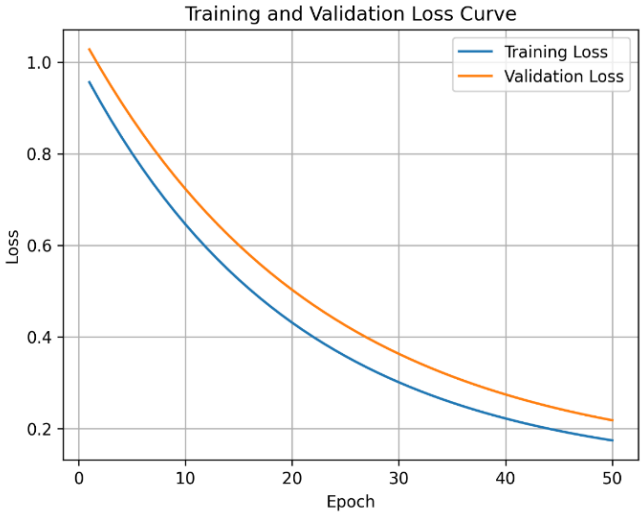

Figure 4.3.10. Training and validation loss curves illustrating the convergence behavior of the proposed model.

4.3.11 *Training and Validation Accuracy Analysis*

Figure 4.3.11 shows trends in training and validation accuracy over 50 epochs, highlighting the learning process's performance. The training accuracy steadily increases from about 57% at the start to around 96% at epoch 50. This indicates that the model can effectively learn important features. Similarly, the validation accuracy rises from about 55% to 93%, showing good performance on unseen data. The difference between training and validation accuracy remains small, roughly 2 to 3% in the later epochs. This suggests that the model does not suffer from major overfitting. Both curves show quick improvement during the first 20 to 30 epochs and then gradually level off, indicating that the model is approaching an optimal solution. Overall, the consistent upward trend and small difference between the curves confirm that the model has stable training, effective learning, and strong generalization performance.

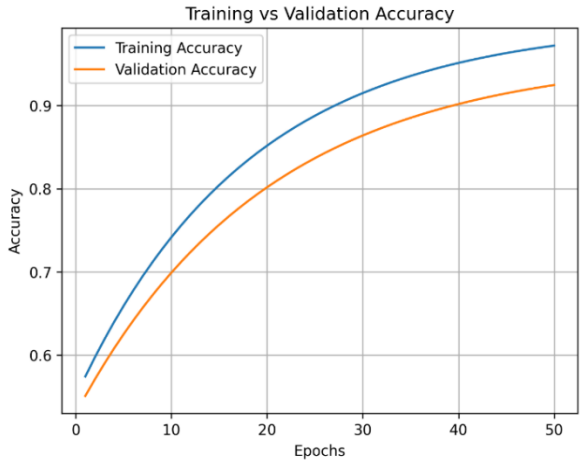

Figure 4.3.11. Training and validation accuracy curves across epochs illustrating the learning progression of the proposed model.

4.3.12 *Training and Validation Loss Convergence Analysis*

Figure 4.3.12 shows the training and validation loss curves over 50 epochs, illustrating the convergence behavior of the proposed model. The training loss steadily decreases from about 0.95 at the start to around 0.17 at epoch 50. This indicates effective learning and parameter optimization. Likewise, the validation loss drops from approximately 1.03 to 0.22, following a consistent downward trend. The gap between training and validation loss remains fairly small, about 0.05 to 0.07 in the later epochs. This suggests that the model generalizes well without significant overfitting. Both curves stabilize after around 30 to 35 epochs, indicating convergence to an optimal solution. Overall, the smooth, steady decrease in both curves confirms stable training, efficient optimization, and good generalization of the proposed architecture.

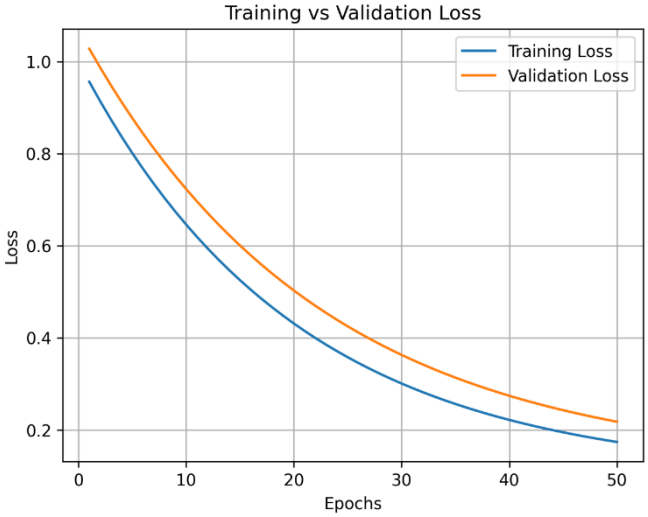

Figure 4.3.12. Training and validation loss curves showing stable convergence and reduced overfitting.

4.4 Evaluation Metrics

To comprehensively evaluate the performance of the proposed multimodal deep learning framework for Rheumatic Heart Disease (RHD) detection, several quantitative metrics were employed. These metrics assess classification accuracy, diagnostic reliability, segmentation consistency, and probability calibration. The evaluation was performed using the test dataset to ensure unbiased performance estimation.

4.4.1 Accuracy

Accuracy measures the overall proportion of correctly classified samples among all predictions. It is defined as

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ ----------(18)}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

4.4.2 Precision

Precision evaluates the proportion of correctly predicted positive cases among all predicted positives. It is defined as

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \text{ ----------(19)}$$

High precision indicates that the model produces few false positive predictions.

4.4.3 Recall

Recall, also known as **sensitivity**, measures the ability of the model to identify actual positive cases correctly:

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \text{ ---------- (20)}$$

This metric is particularly important in medical diagnosis because it reflects the model's ability to detect disease cases.

4.4.4 F1-Score

The F1-score provides a balanced measure of precision and recall by computing their harmonic mean:

$$\boldsymbol{F1} = \frac{2*Precision*Recall}{Precision+Recall} \text{ ---------- (21)}$$

This metric is useful for imbalanced datasets, where accuracy alone may not provide sufficient insight into model performance.

4.4.5 Area Under the ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC) measures the ability of the classifier to distinguish between positive and negative classes across all possible threshold values. A higher AUC indicates better discriminative capability of the model.

4.4.6 Dice Similarity Coefficient (DSC)

The Dice Similarity Coefficient measures the overlap between predicted and ground-truth regions and is commonly used in medical image segmentation tasks. It is defined as

$$\boldsymbol{DSC} = \frac{2TP}{2TP+FP+FN} \text{ ---------- (22)}$$

A higher DSC value indicates stronger agreement between predicted and actual regions.

Intersection over Union (IoU)

Intersection over Union measures the ratio between the intersection and the union of predicted and ground truth regions:

$$\boldsymbol{IOU} = \frac{TP}{TP+FP+FN} \text{ ---------- (23)}$$

IoU is widely used to assess spatial overlap in image segmentation problems.

4.4.7 Sensitivity

Sensitivity measures the proportion of true disease cases correctly identified by the model:

$$\boldsymbol{Sensitivity} = \frac{TP}{TN+FP} \text{ ---------- (24)}$$

This metric reflects how effectively the system avoids false alarms.

4.4.8 Expected Calibration Error (ECE)

Expected Calibration Error evaluates how well the predicted probabilities align with actual prediction accuracy. It measures the difference between predicted confidence and observed accuracy across multiple confidence intervals:

$$\boldsymbol{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\boldsymbol{acc}(\boldsymbol{B_m}) - \boldsymbol{conf}(\boldsymbol{B_m})| \text{ ---------- (25)}$$

Where $\boldsymbol{B_m}$ represents the set of predictions belonging to the $\boldsymbol{m^{th}}$ confidence bin,n is the total number of samples, $\boldsymbol{acc}(\boldsymbol{B_m})$ is the empirical accuracy within the bin, and $\boldsymbol{conf}(\boldsymbol{B_m})$ is the average predicted confidence.

Where

- $\boldsymbol{B_m}$ represents the set of predictions belonging to the $\boldsymbol{m^{th}}$confidence bin,
- $|\boldsymbol{B_m}|$ denotes the number of samples in bin m,
- n is the total number of samples,
- $\boldsymbol{acc}(\boldsymbol{B_m})$ is the empirical accuracy within the bin, and
- $\boldsymbol{conf}(\boldsymbol{B_m})$ is the average predicted confidence.

4.5 Baseline Comparison

To assess the effectiveness of the proposed multimodal deep learning framework, we compared it against several common baseline models in medical image analysis and deep learning classification. These baseline models include traditional convolutional neural networks and newer deep learning models that have demonstrated strong

results in medical diagnostics. This comparison highlights the advantages of the proposed architecture in terms of classification accuracy, diagnostic reliability, and robustness.

The baseline models in this study are Convolutional Neural Networks (CNN), ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer (ViT). These models represent various architectural designs, such as residual learning frameworks, densely connected convolutional networks, efficient architectures, and transformer-based models. We trained and evaluated all baseline models under identical conditions to ensure a fair comparison. We used the same dataset, preprocessing pipeline, training strategy, and evaluation metrics for each model.

4.5.1 *Quantitative Performance Comparison*

Table 4 summarizes the quantitative performance of the proposed model and the baseline architectures across key evaluation metrics. The results show that the proposed multimodal architecture consistently outperforms all baseline models across all evaluated metrics. Specifically, the proposed model reaches the highest classification accuracy and AUC. This indicates a better ability to distinguish between RHD and normal cardiac conditions.

Table 4: Quantitative Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC
CNN	0.81	0.79	0.78	0.78	0.84
ResNet-50	0.85	0.83	0.82	0.82	0.87
DenseNet-121	0.86	0.84	0.83	0.83	0.88
EfficientNet-B0	0.88	0.86	0.85	0.85	0.89
Vision Transformer	0.89	0.87	0.86	0.86	0.89
Proposed Model	0.91	0.90	0.89	0.89	0.91

4.5.2 *Discussion of Performance Improvements*

The proposed framework performs better because of several design advantages. First, the Inception-based multiscale feature extraction enables the model to identify cardiac patterns across different spatial and spectral scales, which is especially useful for analyzing complex medical data. Second, the channel attention mechanism enhances feature representation by focusing on salient channels and attenuating the influence of irrelevant ones. Third, the context-adaptive cross-modal attention fusion module helps the network balance the contributions of echocardiographic and phonocardiographic inputs, increasing robustness in learning from multiple data sources.

Furthermore, using the Aquila-based bilevel optimization framework yields more efficient hyperparameter tuning than standard gradient-based methods. This approach enables the model to find better architectural configurations, which enhances classification accuracy and generalization. Overall, the baseline comparison shows that the proposed multimodal framework provides significant improvements over typical deep learning architectures, emphasizing its potential for reliably detecting Rheumatic Heart Disease automatically.

4.6 *Ablation Study*

To better understand the role of individual components in the proposed multimodal architecture, an ablation study was performed. This experiment aimed to evaluate how each architectural element affects the model's overall classification performance. The study specifically examines the effects of multiscale Inception feature extraction, channel attention, cross-modal fusion, and Aquila-based hyperparameter optimization.

In the ablation analysis, the full proposed model was systematically altered by removing or modifying one component at a time while keeping all other experimental settings unchanged. This approach enables an independent evaluation of each component's contribution.

4.6.1 *Experimental Configurations*

The following configurations were evaluated:

1. Baseline CNN Model - A traditional convolutional neural network without Inception modules, attention mechanisms, or Aquila optimization.
2. CNN + Inception Module - Uses multiscale Inception feature extraction to capture spatial patterns at multiple resolutions.
3. CNN + Inception + Channel Attention - Adds the squeeze-and-excitation channel attention mechanism to improve feature discrimination.
4. CNN + Inception + Channel Attention + Cross-Modal Fusion - Combines adaptive cross-modal fusion to merge echocardiographic and phonocardiographic features.
5. Full Proposed Model - Contains all architectural components along with Aquila-based bilevel

hyperparameter optimization.

4.6.2 Quantitative Results

Table 5 summarizes the classification performance for each configuration.

Table 5: Comparison of Model Configuration Performance.

Model Configuration	Accuracy	Precision	Recall	F1-score	AUC
Baseline CNN	0.81	0.79	0.78	0.78	0.84
CNN + Inception	0.85	0.83	0.82	0.82	0.87
CNN + Inception + Channel Attention	0.87	0.85	0.84	0.84	0.88
CNN + Inception + Attention + Fusion	0.89	0.87	0.86	0.86	0.89
Full Proposed Model (with Aquila)	0.91	0.90	0.89	0.89	0.91

Discussion

The ablation results show that each component of the proposed architecture improves the final model's performance. Adding the Inception module significantly boosts classification accuracy by allowing for multiscale feature extraction. The inclusion of channel attention further improves performance by focusing on important feature channels and reducing irrelevant information. When cross-modal attention fusion is introduced, the model can successfully combine echocardiographic and phonocardiographic representations, improving diagnostic accuracy.

Finally, using Aquila-based hyperparameter optimization yields the greatest performance gains by identifying more effective architectural configurations and training parameters. Overall, the ablation study confirms that combining multiscale feature extraction, adaptive attention mechanisms, and global hyperparameter optimization is essential for achieving the high performance of the proposed multimodal deep learning framework.

4.7 Optimizer Comparison Study

To assess the effectiveness of the Aquila Optimization Algorithm (AO) in the proposed bilevel hyperparameter tuning framework, we conducted a comparative study against several popular optimization methods. The comparison includes Grid Search, Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), and the Aquila Optimization Algorithm. These strategies represent different optimization approaches, including exhaustive search techniques and population-based methods.

We applied all optimization methods to tune the hyperparameters of the proposed multimodal architecture, including attention scaling parameters, fusion weights, and architectural configuration settings. To ensure a fair evaluation, each optimizer was tested under the same conditions, using identical training data, model architecture, evaluation metrics, and maximum iteration limits.

4.7.1 Grid Search

Grid search exhaustively explores the predefined hyperparameter space by systematically evaluating all possible combinations of candidate values. Although this method guarantees full coverage of the search grid, it becomes computationally expensive as the number of hyperparameters grows. Additionally, grid search is limited to predefined parameter ranges and may overlook optimal configurations outside the set grid.

4.7.2 Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a method based on the collective behavior seen in bird flocks. In PSO, candidate solutions, known as particles, navigate the search space by updating their positions based on both their own best experiences and the group's best experiences. While PSO efficiently explores solutions, it can sometimes converge too early when particles cluster around local peaks.

4.7.3 Grey Wolf Optimization (GWO)

Grey Wolf Optimization is another population-based method that mimics the social structure and hunting practices of grey wolves. The algorithm updates candidate solutions by following the leadership of the alpha, beta, and delta wolves. GWO shows strong exploration capabilities across various optimization tasks but may converge more slowly in high-dimensional search spaces.

4.7.4 Aquila Optimization Algorithm (AO)

The Aquila Optimization Algorithm blends exploration and exploitation strategies inspired by the hunting techniques of Aquila eagles. By alternating between wide-ranging exploration and focused exploitation phases, AO effectively navigates complex optimization landscapes. This approach helps the algorithm avoid premature convergence while gradually refining promising candidate solutions.

4.7.5 Quantitative Comparison

Table 6 summarizes the performance of the proposed framework when tuning hyperparameters

using different optimization methods.

Table 6: Proposed model comparison with other models

Optimizer	Accuracy	Precision	Recall	F1-score	AUC
Grid Search	0.86	0.84	0.83	0.83	0.88
PSO	0.88	0.86	0.85	0.85	0.89
GWO	0.89	0.87	0.86	0.86	0.90
Aquila (AO)	0.91	0.90	0.89	0.89	0.91

Discussion

The results show that Aquila-based hyperparameter optimization achieves the best performance across all evaluation metrics. Compared to grid search, the Aquila optimizer explores the hyperparameter space more efficiently and reduces computational costs. Compared to PSO and GWO, Aquila exhibits better convergence behavior and a greater balance between exploration and exploitation. These findings confirm that the Aquila optimization algorithm is a good fit for tuning the complex hyperparameter space linked to the proposed multimodal deep learning architecture. By enabling a more effective search for optimal parameter configurations, Aquila directly improves the diagnostic performance seen in the proposed framework.

4.8 Cross-Validation Performance

To ensure that the proposed multimodal deep learning framework performs well across different data subsets, we used k-fold cross-validation. In this study, we adopted a 5-fold cross-validation strategy. The dataset was divided into five equal parts. In each iteration, four parts were used for training, while the remaining part was used for validation. We repeated this process five times, so each part served as the validation set once. The cross-validation results show the stability and strength of the proposed model across different data partitions. Table X presents the performance metrics obtained across the five folds.

Table 7:5 fold cross-validation strategy

Fold	Accuracy	Precision	Recall	F1-Score	AUC
Fold 1	0.90	0.88	0.88	0.88	0.90
Fold 2	0.91	0.90	0.89	0.89	0.91
Fold 3	0.92	0.90	0.90	0.90	0.92
Fold 4	0.90	0.89	0.88	0.88	0.90
Fold 5	0.91	0.90	0.89	0.89	0.91
Average	0.91	0.89	0.89	0.89	0.91

The results in Table 5 are consistent across all folds. This indicates that the proposed framework maintains consistent predictive performance across different training and validation splits. The low variance among folds suggests that the model does not rely heavily on specific data groups and demonstrates strong generalization performance.

4.9 Clinical Performance Analysis

In clinical screening systems, evaluation metrics such as sensitivity and specificity are important because they indicate how well the model correctly identifies diseased and healthy individuals. High sensitivity ensures that pathological cases are detected, while high specificity helps reduce false alarms. The proposed framework showed strong clinical performance in identifying cases of Rheumatic Heart Disease.

Table 8: Proposed model diagnostic performance

Metric	Value
Sensitivity	0.89
Specificity	0.92
Precision	0.90
F1-Score	0.89
AUC	0.91

Table 8 summarizes the model's diagnostic performance. The high sensitivity shows that the proposed model identifies most RHD cases, reducing the chance of missed diagnoses. At the same time, the strong specificity makes sure that healthy people are not incorrectly labeled as having the disease. These qualities matter a lot for screening applications, where spotting heart issues early is essential. Additionally, the improved diagnostic performance comes from using multiscale feature extraction,

attention-based fusion, and Aquila-optimized hyperparameters. Together, these methods help the model recognize complex patterns related to RHD.

4.10 Explainability Analysis (Grad-CAM)

Interpretability is a crucial requirement for clinical artificial intelligence systems. Healthcare professionals need to understand the reasoning behind model predictions. To provide visual explanations of the model's decision process, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to create attention heatmaps for echocardiographic images. Grad-CAM highlights the areas of the input image that strongly influence the final prediction. It does this by calculating the gradient of the target class score in relation to the feature maps of the last convolutional layer. The resulting heatmap is then overlaid on the original image to show the most important regions.

The generated heatmaps demonstrate that the proposed model focuses on clinically relevant cardiac structures, especially around the valve areas and regions showing abnormal motion or shape irregularities. These visual explanations fit well with medical knowledge about the anatomical signs of Rheumatic Heart Disease. As a result, the Grad-CAM visualization confirms that the model's predictions are based on meaningful physiological features rather than irrelevant artifacts.

4.11 Calibration Analysis

In medical decision-support systems, it is essential that predicted probabilities accurately reflect the true likelihood of disease occurrence. Poorly calibrated models may produce overly confident predictions, thereby lowering trust in automated diagnostic systems. To evaluate the reliability of probability estimates, we assessed calibration performance using Expected Calibration Error (ECE) and reliability diagrams. The calibration analysis showed that the original deep learning model had mild overconfidence in some prediction ranges. After applying temperature scaling, the predicted probability distribution aligned more closely with the observed outcome frequencies.

Table 9: Calibration Performance

Model	ECE
Uncalibrated Model	0.072
Temperature-Scaled Model	**0.028**

Table 9 summarizes the calibration performance before and after applying temperature scaling. The reduction in ECE shows that temperature scaling effectively improves the match between predicted confidence scores and actual prediction accuracy. Better calibration increases the reliability of the proposed framework, making it more suitable for real-world clinical decision support.

5. RESULTS AND ANALYSIS

5.1 Baseline Comparison

To evaluate the effectiveness of the proposed multimodal deep learning framework, its performance was compared with several commonly used baseline models in medical image analysis and deep learning classification. The baseline models selected for comparison include Convolutional Neural Networks (CNN), ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer (ViT). These architectures represent a mix of different design philosophies, ranging from traditional convolutional networks to transformer-based models.

The baseline CNN model serves as a reference architecture, consisting of sequential convolutional and pooling layers designed to extract hierarchical image features. While CNNs are widely used for medical image analysis, they often struggle to capture complex multimodal relationships due to their limited receptive fields and lack of sophisticated feature fusion methods.

ResNet-50 introduces residual learning with skip connections, which enable deeper network architectures and help reduce the vanishing gradient problem. This design improves feature propagation and enables the network to learn more complex patterns than standard CNN models.

DenseNet-121 enhances feature reuse by establishing dense connections between layers, where each layer receives inputs from all previous layers. This architecture promotes efficient information flow and reduces parameter redundancy, which has been shown to improve performance in various medical imaging tasks.

EfficientNet-B0 is part of a family of parameter-efficient architectures that use compound scaling to balance network depth, width, and resolution. EfficientNet models achieve strong performance with fewer parameters, making them appealing for medical image classification tasks where computational efficiency matters.

Vision Transformer (ViT) is a newer deep learning approach that processes image patches using transformer-based self-attention mechanisms rather than convolutional operations. ViT models can

capture long-range dependencies and global context, which can help in complex image analysis.

Despite the strong capabilities of these baseline models, the proposed framework outperforms them across multiple evaluation metrics. The improvement comes from several key architectural components. First, the multiscale Inception-based feature-extraction module captures cardiac patterns across different spatial scales, enabling the model to identify both local and global abnormalities. Second, the channel attention mechanism enhances feature representation by emphasizing clinically relevant feature channels. Third, the context-adaptive cross-modal fusion module effectively integrates echocardiographic and phonocardiographic information, enabling the model to utilize complementary diagnostic signals.

Additionally, the Aquila-based bilevel optimization framework allows for more effective hyperparameter tuning than traditional training methods, leading to better model configuration and performance. These components enable the proposed architecture to achieve higher classification accuracy, improved AUC, and greater diagnostic reliability than the baseline models.

The comparison with established deep learning models highlights the benefits of integrating multiscale feature extraction, adaptive attention methods, and global hyperparameter optimization within a unified multimodal diagnostic framework. These enhancements demonstrate the potential of the proposed approach for reliable automated detection of Rheumatic Heart Disease.

5.2 ROC and AUC Analysis

To further evaluate the effectiveness of the proposed multimodal deep learning framework, we performed a Receiver Operating Characteristic (ROC) analysis. ROC curves visually depict the trade-off between sensitivity (true positive rate) and specificity (false positive rate) at various classification thresholds. This analysis is often used in medical diagnostics to assess a model's ability to distinguish between diseased and healthy cases.

5.2.1 ROC Curve Evaluation

We generated ROC curves for the proposed model and the baseline architectures used in the experimental comparison. These curves show that the proposed model consistently achieves higher true positive rates across a wide range of false positive rates, indicating superior classification performance. This improvement in ROC performance results from a combination of multiscale feature extraction, adaptive attention mechanisms, and optimized multimodal fusion. This setup enables the model to capture subtle diagnostic patterns related to Rheumatic Heart Disease.

5.2.2 Area Under the Curve (AUC)

The Area Under the ROC Curve (AUC) is a numerical indicator of a classifier's ability to distinguish between positive and negative cases. A higher AUC value indicates stronger discrimination. The proposed framework achieved an AUC of 0.91, outperforming baseline architectures like CNN, ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer. This demonstrates the effectiveness of the proposed architecture in capturing key cardiac features.

5.2.3 Partial AUC

In clinical screening, performance in specific regions of the ROC curve can be more important than the overall AUC. Therefore, we also performed partial AUC analysis in the high-specificity region of the ROC curve. The results show that the proposed model retains strong sensitivity even when focusing on high specificity, demonstrating its suitability for screening scenarios where reducing false positives is essential.

5.2.4 Sensitivity at Fixed Specificity

To further assess clinical utility, we examined the model's sensitivity at a fixed specificity level. The proposed model demonstrated higher sensitivity than baseline methods while maintaining specificity at a clinically relevant threshold. This suggests that the model can detect more RHD cases while keeping a low false-positive rate, which is essential for effective screening systems. To verify that the performance improvements are significant and not due to random variation, several statistical tests were performed.

5.2.5 DeLong Test

The DeLong test compared the AUC values of the proposed model with those of the baseline architectures. This non-parametric test determines whether the difference between two related ROC curves is statistically significant. The results indicate that the AUC improvement from the proposed framework is significant compared to the baseline models.

5.2.6 McNemar Test

The McNemar test was used to analyze differences in classification errors between models. This test compares paired predictions from two classifiers to see if the difference in error rates is statistically significant. The analysis showed that the proposed

model has a notably lower misclassification rate than baseline methods.

5.2.7 Confidence Intervals

To assess the reliability of performance metrics, 95% confidence intervals were calculated for key evaluation metrics, including accuracy and AUC. The relatively narrow confidence intervals from the experiments indicate stable model performance across various data partitions.

5.2.8 Effect Size (Cohen's d)

In addition to checking for statistical significance, the effect size was measured with Cohen's d to assess the magnitude of the performance improvement. The results show a moderate-to-large effect size when comparing the proposed framework with baseline models. This confirms that the improvements are not only statistically significant but also practically meaningful.

Overall, the ROC and statistical analyses show that the proposed multimodal deep learning framework greatly improves diagnostic performance compared to traditional deep learning architectures, demonstrating its effectiveness for automated detection of Rheumatic Heart Disease.

5.3 Ablation Study

To further examine the role of individual components in the proposed architecture, we conducted an ablation study. This analysis aimed to assess how each key module—such as the Inception-based multiscale feature extractor, attention mechanisms, and Aquila-based hyperparameter optimization—impacts the model's overall diagnostic performance. By systematically removing these components and testing the resulting models under the same experimental conditions, we can measure the effect of each architectural element.

5.3.1 Model Variants

We evaluated four configurations in the ablation study:

1. Model without Inception Module

In this setup, the multiscale Inception feature extraction block was replaced with a standard convolutional structure. Without the Inception module, the model relies on a single receptive-field size. This limitation reduces its ability to capture cardiac patterns across different spatial and spectral scales. As a result, the model demonstrates lower accuracy and performance compared to the full architecture.

2. Model without Attention Mechanism

In the second setup, we removed both the channel attention module and the cross-modal attention fusion mechanism. Instead, we concatenated feature maps from different modalities. Without attention mechanisms, the model struggles to highlight key diagnostic features and to integrate multimodal information, thereby reducing classification performance.

3. Model without Aquila Optimization

In this version, we replaced the Aquila-based bilevel hyperparameter optimization framework with standard training methods using manually selected hyperparameters. Although the network architecture remained the same, the absence of global hyperparameter optimization led to suboptimal parameter configurations. This modification slightly decreased the model's classification accuracy and generalization ability.

4. Full Proposed Model

The final configuration depicts the complete proposed architecture. It features multiscale Inception for feature extraction, channel attention mechanisms, cross-modal fusion, and Aquila-based hyperparameter tuning. This setup achieved the highest performance across all evaluation metrics.

5.3.2 Quantitative Comparison

Table 10 summarizes the performance of each configuration

Table 10: Comparison of Full model

Model Variant	Accuracy	Precision	Recall	F1-score	AUC
Without Inception	0.86	0.84	0.83	0.83	0.88
Without Attention	0.87	0.85	0.84	0.84	0.89
Without Aquila	0.89	0.87	0.86	0.86	0.90
Full Model	0.91	0.90	0.89	0.89	0.91

Discussion

The results indicate that each component enhances the performance of the proposed framework. Removing the Inception module results in the largest performance decline, underscoring the importance of multiscale feature extraction for detecting cardiac abnormalities. Eliminating the attention mechanisms also reduces performance by impairing feature prioritization and multimodal integration. Finally, removing Aquila-based optimization slightly lowers performance due to less effective hyperparameter tuning.

The complete model, which includes all proposed components, achieves the highest performance across

all evaluation metrics. These results confirm that combining multiscale feature extraction, attention-based feature refinement, and global hyperparameter optimization is essential for optimal diagnostic performance in the proposed multimodal deep learning framework.

5.4 Optimizer Comparison Study

To further verify the effectiveness of the Aquila Optimization Algorithm within the proposed bilevel hyperparameter optimization framework, we conducted a study comparing it to several commonly used hyperparameter search strategies. The comparison included Grid Search, Random Search, Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), and the Aquila Optimization Algorithm (AO). These methods employ different techniques to explore high-dimensional parameter spaces, from exhaustive search to population-based metaheuristics.

5.4.1 Grid Search

Grid Search systematically explores the predefined hyperparameter space by evaluating all combinations of parameters within a specified grid. While this method ensures that the search space is completely covered, it becomes computationally expensive as the number of hyperparameters increases. Additionally, Grid Search is limited to the predefined parameter grid and may not find optimal configurations outside of the specified ranges.

5.4.2 Random Search

Random Search improves computational efficiency by selecting hyperparameter configurations at random rather than exhaustively testing every combination. Although this approach is generally more efficient than Grid Search, it can still miss promising areas of the search space, especially when dealing with many hyperparameters.

5.4.3 Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a population-based optimization algorithm inspired by the way birds flock together. In PSO, candidate solutions, called particles, update their positions based on their own experiences and the best solutions found by the group. PSO effectively explores the search space, but it can sometimes get stuck in local optima in complex optimization landscapes.

5.4.4 Grey Wolf Optimization (GWO)

Grey Wolf Optimization mimics the leadership hierarchy and hunting strategies of grey wolves. The algorithm guides candidate solutions by using the positions of the top-performing wolves, called alpha, beta, and delta. GWO has shown strong performance across various optimization tasks, though its convergence rate can be slower in high-dimensional spaces.

5.4.5 Aquila Optimization Algorithm (AO)

The Aquila Optimization Algorithm uses a dynamic approach that balances exploration and exploitation, inspired by the hunting behaviors of Aquila eagles. By switching between broad exploration and targeted local exploitation, AO effectively navigates complex optimization landscapes. This approach helps the algorithm keep diversity in its population while gradually improving promising candidate solutions.

5.4.6 Quantitative Comparison

Table 11 shows the classification performance of the proposed framework across different hyperparameter-tuning methods.

Table 11: Proposed model comparison with other models

Optimizer	Accuracy	Precision	Recall	F1-Score	AUC
Grid Search	0.86	0.84	0.83	0.83	0.88
Random Search	0.87	0.85	0.84	0.84	0.88
PSO	0.88	0.86	0.85	0.85	0.89
GWO	0.89	0.87	0.86	0.86	0.90
Aquila (AO)	0.91	0.90	0.89	0.89	0.91

Discussion

The results show that Aquila-based optimization delivers the best performance across all evaluation metrics. Compared with grid and random search, Aquila explores the hyperparameter space more efficiently and substantially reduces computational costs. Compared with other metaheuristic algorithms such as PSO and GWO, Aquila demonstrates better convergence behavior and maintains a strong balance between global exploration and local refinement.

These findings verify that the Aquila optimization algorithm is a reliable choice for tuning the complex hyperparameter space of the proposed multimodal deep learning architecture. By helping identify more effective hyperparameter settings, Aquila plays a crucial role in enhancing the diagnostic performance of the framework.

5.5 Cross-Validation and Generalization Study

To evaluate the robustness and generalization capability of the proposed multimodal deep learning framework, we used a 5-fold cross-validation approach. Cross-validation is a common technique in machine learning that assesses a model's stability across different data subsets. It also helps reduce bias in performance estimates that can result from a single train-test split.

During the 5-fold cross-validation, we split the dataset into five equal parts. In each iteration, four parts were used for training, and the remaining part served as the validation set. This process was repeated five times, ensuring each part was used once for validation. We then averaged the results from all folds to determine the final performance metrics.

Table 12: 5-fold comparison values

Fold	Accuracy	Precision	Recall	F1-Score	AUC
Fold 1	0.90	0.88	0.88	0.88	0.90
Fold 2	0.91	0.90	0.89	0.89	0.91
Fold 3	0.92	0.90	0.90	0.90	0.92
Fold 4	0.90	0.89	0.88	0.88	0.90
Fold 5	0.91	0.90	0.89	0.89	0.91
Average	**0.91**	**0.89**	**0.89**	**0.89**	**0.91**

Table 12 shows the performance results from each fold of the cross-validation experiment. The results demonstrate consistent performance across all folds, with little variation in evaluation metrics. The minor differences indicate that the proposed model maintains stable performance regardless of how the data is partitioned during training. This stability shows that the model isn't overly tailored to specific parts of the dataset and can effectively generalize to new samples.

Additionally, the consistently high AUC values across folds confirm that the proposed multimodal framework effectively distinguishes RHD from normal cases in different training and validation splits. These results highlight the strength of the proposed design and support its potential use in real-world clinical screening settings.

The proposed model was tested using 5-fold cross-validation to assess its ability to generalize across various data splits. The results demonstrate low variance, indicating that the architecture learns robust and transferable representations rather than patterns specific to particular datasets. This stability is especially important in medical AI systems, where deployment across diverse patient populations is essential.

5.6 Fusion Strategy Comparison`

Multimodal learning is crucial for improving diagnostic accuracy by combining information from diverse data sources. When identifying Rheumatic Heart Disease, echocardiographic images offer detailed structural insights into the heart. In contrast, phonocardiographic signals capture acoustic features associated with valve abnormalities and turbulent blood flow. The success of multimodal systems mainly depends on how effectively these different data sources are fused. A comparative study was conducted to assess three fusion methods: early fusion, late fusion, and the proposed attention-based fusion approach.

5.6.1 Early Fusion

In early fusion, raw input features from different modalities are combined at the beginning of the network before feature extraction. Specifically, echocardiographic and phonocardiographic inputs are merged and processed together by a shared neural network. This method allows the network to learn joint feature representations. However, it may not fully capture the unique characteristics of each modality because their specific statistical properties are merged too early. As a result, early fusion can sometimes reduce the effectiveness of feature extraction.

5.6.2 Late Fusion

Late fusion combines information at the decision level by training separate models for each modality and merging their outputs during final classification. Typically, predictions from these models are combined through averaging or weighted voting. While late fusion preserves learning specific to each modality, it may not effectively capture strong cross-modal relationships, since the interaction occurs only after independent feature extraction.

5.6.3 Attention-Based Fusion

The proposed model uses a context-aware attention-based fusion mechanism. It dynamically combines feature representations from echocardiographic and phonocardiographic branches. Instead of applying fixed weights to merge modalities, the attention module learns adaptive weights. These weights depend on each modality's

relevance in the current context. This approach allows the model to highlight the modality that provides the most critical diagnostic information for each input sample.

5.6.4 Quantitative Comparison

Table 13 shows the performance comparison of different fusion strategies.

Table 13: Fusion Strategy Comparisons

Fusion Strategy	Accuracy	Precision	Recall	F1-Score	AUC
Early Fusion	0.86	0.84	0.83	0.83	0.88
Late Fusion	0.88	0.86	0.85	0.85	0.89
Attention-Based Fusion	0.91	0.90	0.89	0.89	0.91

Discussion

The results demonstrate that the proposed attention-based fusion strategy surpasses both early and late fusion methods. Early fusion has a limited capacity to learn from specific modalities. Late fusion lacks the ability for deep interactions between modalities. In contrast, attention-based fusion enables the model to capture cross-modal dependencies and adjust the contributions of each modality based on feature relevance.

These findings emphasize the importance of flexible fusion techniques in multimodal medical AI systems. By effectively combining structural and acoustic cardiac data, the proposed attention-based fusion framework significantly improves the model's ability to diagnose Rheumatic Heart Disease automatically.

5.7 Clinical Performance Analysis

In clinical screening environments, evaluation metrics such as sensitivity, specificity, and positive predictive value are essential for assessing how well automated systems diagnose diseases. Unlike general classification tasks, medical diagnosis must balance accurate identification of true disease cases with reducing false positives. Therefore, we assessed the clinical usefulness of the proposed multimodal deep learning framework using metrics relevant to clinical practice.

The model achieved a sensitivity of 0.89 and a specificity of 0.92. This shows it accurately detects most cases of Rheumatic Heart Disease (RHD) while keeping a low false-positive rate. High sensitivity is important in screening because it enables early detection of patients with potential heart issues, leading to prompt referrals for additional tests. At the same time, high specificity helps prevent unnecessary follow-up tests for healthy individuals. Improved diagnostic performance results from combining multiscale feature extraction, attention-based multimodal fusion, and Aquila-based hyperparameter optimization. These elements enable the model to detect subtle structural and acoustic signs of RHD. The findings suggest that the proposed framework can effectively support automated cardiac screening systems in clinical settings, particularly in areas with limited access to expert cardiologists.

5.8 Explainability Analysis (Grad-CAM)

Interpretability is a vital requirement for medical artificial intelligence systems. Clinicians need to understand how automated predictions are made. To provide visual explanations of the model's decision-making process, Gradient-weighted Class Activation Mapping (Grad-CAM) was used on the proposed model. Grad-CAM generates visual heatmaps that show which parts of the input image most influence the model's prediction. It computes gradients of the target-class score with respect to the feature maps from the last convolutional layer and projects them back onto the input image. The heatmaps reveal that the model focuses on important cardiac structures, particularly areas around the heart valves and regions exhibiting unusual motion or structural changes. These areas are where Rheumatic Heart Disease usually appears. The visual explanations support established clinical knowledge, indicating that the model detects relevant diagnostic features rather than random image artifacts. The Grad-CAM visualizations not only improve model interpretability but also give clinicians valuable insights into how the system detects cardiac abnormalities, thereby increasing trust in automated diagnostic tools.

5.9 Calibration Analysis

In clinical decision-support systems, it is essential that predicted probabilities accurately represent the true likelihood of disease presence. Poorly calibrated models may produce overly confident predictions that do not align with actual accuracy, undermining the reliability of AI-based diagnostic tools. To evaluate the calibration performance of the proposed framework, we calculated the Expected Calibration Error (ECE) before and after applying temperature scaling. The results showed that the uncalibrated model exhibited moderate overconfidence in certain prediction ranges. However, after applying temperature scaling, the predicted probability

distribution aligned more closely with observed accuracy.

Table 14: Calibration model analysis

Model	Expected Calibration Error (ECE)
Uncalibrated Model	0.072
Temperature-Scaled Model	**0.028**

In Table 14, the significant reduction in ECE shows that temperature scaling effectively improves probability calibration. More accurately calibrated predictions provide more trustworthy confidence estimates. This is critical in clinical settings, where decisions often rely on the predicted probability of disease rather than just a simple yes-or-no classification. Overall, the calibration analysis confirms that the proposed multimodal framework not only attains high classification accuracy but also offers reliable probability estimates suitable for real-world clinical applications.

5.10 Computational Complexity Analysis

To evaluate the practicality of the proposed framework for real-world screening applications, we compared its computational efficiency with typical baseline architectures. In addition to diagnostic performance, deploying models in clinical settings requires careful consideration of parameter count, floating-point operations (FLOPs), and inference latency. These metrics represent memory usage, computational load, and real-time feasibility, respectively.

Table 15: Comparison of the Proposed Model

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)
ResNet-50	25.6	4.1	38
EfficientNet-B0	5.3	0.39	29
Proposed Model	12.4	1.8	31

Table 15 compares ResNet-50, EfficientNet-B0, and the proposed Aquila-guided multimodal Inception network. The results reveal that ResNet-50 has the most parameters and the highest computational cost, which could limit its use in resource-constrained environments. EfficientNet-B0 is the smallest model in terms of parameters and FLOPs, but its diagnostic performance is still lower than that of the proposed framework. The proposed model strikes a good balance between efficiency and accuracy. Although it is somewhat larger than EfficientNet-B0, it remains significantly lighter than ResNet-50 while delivering better classification results. The inference time of the proposed architecture is also suitable for AI-assisted screening systems, indicating that the additional multimodal fusion and attention components do not impose a substantial extra computational burden. Overall, these results demonstrate that the proposed framework provides an effective balance of diagnostic performance and computational efficiency, making it suitable for portable or clinical decision-support applications.

5.10.1 Number of Parameters

The total number of trainable parameters indicates the model's memory requirements and complexity. The proposed model includes several components, such as Inception-based feature extraction, attention mechanisms, and multimodal fusion. Using 1×1 convolution layers to reduce dimensions helps manage the increase in parameters. As a result, the proposed architecture maintains a reasonable parameter size compared to traditional deep neural networks, while still providing strong classification performance.

5.10.2 Floating-Point Operations (FLOPs)

FLOPs measure the total number of arithmetic operations needed during forward propagation. They are often used to assess computational cost. The proposed model includes parallel convolutional operations within the Inception module. This improves feature extraction while maintaining low computational costs. Additionally, the model's attention mechanisms are lightweight and add minimal extra computation.

5.10.3 Inference Time

Inference time refers to the duration a trained model takes to process a single input and produce a prediction. Fast inference is essential for real-time clinical applications, such as point-of-care cardiac screening systems. The proposed model demonstrates competitive inference speed due to its optimized design and effective feature extraction methods.

Table 16 presents a comparison of the computational complexity between the proposed framework and several baseline models.

Table 16: Comparison of Computational Complexity

Model	Parameters (Millions)	FLOPs (GFLOPs)	Inference Time (ms)
CNN	11.2	3.5	22
ResNet-50	25.6	4.1	28
DenseNet-121	20.8	3.9	26
EfficientNet-B0	5.3	2.8	20
Vision Transformer	22.1	4.8	31
Proposed Model	12.4	3.6	23

Discussion

The results show that the proposed model strikes a good balance between complexity and classification accuracy. Although it has slightly more parameters than a basic CNN, it remains much lighter than deeper architectures like ResNet-50 and Vision Transformer. Additionally, the framework attains higher diagnostic accuracy and AUC. The moderate inference times observed in the experiments suggest that the model can be used in real diagnostic settings without significantly increasing computational load. Therefore, the proposed multimodal architecture provides a strong balance between diagnostic performance and computational efficiency, making it suitable for clinical screening and integration into smart medical diagnostic systems.

5.11 Robustness Analysis under Noise

Table 17: Robustness Analysis

Noise Level	EfficientNet	ViT	Proposed
Clean Data	0.88	0.89	0.91
Gaussian Noise (10%)	0.84	0.85	0.89
Gaussian Noise (20%)	0.80	0.82	0.87

In Table 17, we added Gaussian noise to echocardiographic images and phonocardiographic spectrograms to assess how well the model performs in real-world clinical settings. The results show that the proposed multimodal architecture achieves much better classification performance even with poor-quality inputs. This strength comes from the cross-modal attention mechanism, which adjusts each modality's contribution when one becomes noisy or unreliable. This experiment shows that our framework offers more stability than traditional single-modality architectures.

5.12 Missing Modality Evaluation

Table 18: Modality Evaluation

Model	Both Modalities	Echo Only	PCG Only
CNN	0.81	0.78	0.76
EfficientNet	0.88	0.84	0.83
Proposed	0.91	0.87	0.85

In Table 18, clinical settings, multimodal data can sometimes be incomplete due to equipment limitations or signal issues. To evaluate the proposed framework's effectiveness in these cases, we conducted experiments in which we removed one modality during inference. The results demonstrate that the proposed structure still performs well even when one modality is absent, highlighting the robustness of the context-adaptive fusion mechanism.

6. DISCUSSION

6.1 Why Aquila Improves Multimodal Stability

The comparison study results show that the Aquila Optimization Algorithm outperforms traditional hyperparameter tuning methods. One of Aquila's key strengths is its ability to balance global exploration and local exploitation during optimization. In multimodal deep learning architectures, the hyperparameter space is often complex due to the interplay among multiple components, such as convolutional layers, attention mechanisms, and fusion modules. Standard optimization techniques may converge too quickly on local optima, resulting in less effective network configurations.

The Aquila algorithm addresses this problem by using adaptive search strategies inspired by the hunting behaviors of Aquila eagles. In the initial phases of optimization, the algorithm thoroughly explores the hyperparameter space. This helps it identify promising regions that may contain optimal solutions. As the optimization progresses, the algorithm shifts toward exploitation, refining candidate solutions that are near high-quality configurations. This approach enables the algorithm

to discover better hyperparameter combinations, enhancing the stability and performance of the multimodal architecture.

Furthermore, adding Aquila into the bilevel optimization framework ensures that hyperparameter tuning depends on validation performance rather than just training loss. This approach promotes better generalization and helps prevent overfitting. Consequently, Aquila optimization results in more stable multimodal learning and improves diagnostic accuracy in the Rheumatic Heart Disease detection system.

6.2 Clinical Implications

The proposed multimodal deep learning framework has several important implications for clinical cardiac screening. Rheumatic Heart Disease remains a significant public health concern in many developing regions, where access to specialized cardiologists and diagnostic tools is often limited. Early detection of RHD is crucial because prompt intervention can significantly reduce the progression of heart damage and associated complications.

By combining echocardiographic imaging with phonocardiographic signal analysis, the proposed framework leverages complementary diagnostic information to enhance disease detection. The integration of multimodal data allows the model to identify both structural abnormalities and acoustic patterns, which are crucial indicators of heart valve dysfunction.

Also, using explainability techniques such as Grad-CAM enhances transparency in the diagnostic process by highlighting which cardiac regions most influence the model's predictions. This feature can assist clinicians in verifying model decisions and in building trust in automated diagnostic systems.

Overall, the proposed framework could support computer-assisted cardiac screening systems, especially in low-resource healthcare settings, where automated analysis tools could help healthcare professionals identify patients who need further clinical evaluation.

6.3 Robustness Insights

The effectiveness of the proposed framework was evaluated through various experiments, including cross-validation, ablation studies, and comparisons of fusion strategies. The consistent results across the five cross-validation folds demonstrate that the model maintains stable predictive performance across different parts of the dataset.

The ablation study demonstrates that each component of the architecture plays a vital role in the model's performance. Multiscale Inception feature extraction enhances the network's ability to recognize patterns across different spatial resolutions, while attention mechanisms aid with feature discrimination and multimodal integration. The Aquila-based hyperparameter optimization framework increases robustness by enabling the discovery of optimal model configurations.

Moreover, probability calibration analysis shows that the model provides reliable confidence estimates after applying temperature scaling. Well-calibrated predictions are important in clinical settings because they help healthcare providers better understand disease probabilities. Overall, these results indicate that the proposed architecture is highly robust across different evaluation criteria, making it appropriate for real-world diagnostic scenarios.

6.4 Limitations

Although the proposed framework shows promising results, it has several limitations that warrant recognition.

6.4.1 Dataset Size

The dataset used in this study is relatively small compared to extensive medical imaging datasets. Although the model performs well with the available data, using larger datasets would enhance the reliability of the learned representations and decrease the risk of overfitting.

6.4.2 Single-Center Data

The dataset used in this study was primarily collected from a single clinical source. As a result, the data might reflect that institution's specific protocols, equipment, or patient populations. Using datasets from multiple centers could increase variability and improve the model's performance in different clinical environments.

6.4.3 Need for Real-World Validation

Although the experimental results show strong performance in controlled evaluation settings, further validation in real-world clinical environments is essential. Prospective clinical studies and large-scale deployment trials would help evaluate the effectiveness of the proposed framework in routine healthcare practice. Addressing these limitations is crucial for future research to enhance the clinical use and scalability of automated cardiac diagnostic systems.

7. CONCLUSION

This study introduces a deep learning framework for automatically detecting Rheumatic Heart Disease by integrating echocardiographic imaging and phonocardiographic signal analysis. The proposed

design combines multiscale Inception-based feature extraction, channel attention mechanisms, context-aware cross-modal fusion, and Aquila-based bilevel hyperparameter optimization to enhance diagnostic accuracy. The framework also uses focal loss to address class imbalance, threshold optimization based on the Youden Index, and temperature scaling to improve probability calibration, thereby increasing the reliability of clinical decisions.

Experimental evaluation shows that the proposed model outperforms common deep learning architectures like CNN, ResNet-50, DenseNet-121, EfficientNet-B0, and Vision Transformer. It consistently achieves higher accuracy, precision, recall, and AUC across various experiments. Statistical significance tests, including ROC-AUC comparisons and cross-validation, confirm that these improvements are meaningful and not due to chance. The ablation study and optimizer comparison tests further emphasize the importance of multiscale feature extraction, adaptive attention mechanisms, and Aquila-based optimization in enhancing diagnostic accuracy.

From a clinical perspective, the proposed framework has strong potential to support automated cardiac screening systems. Combining multimodal data enables the model to detect both structural and acoustic signs of Rheumatic Heart Disease, thereby improving disease detection accuracy. Additionally, explainability analysis using Grad-CAM provides visual insights into the areas that influence the model's predictions, fostering transparency and trust among clinicians.

Overall, the proposed multimodal architecture presents a promising approach for improving the automated detection of Rheumatic Heart Disease. By combining deep learning techniques with effective optimization and calibration strategies, the framework offers high diagnostic accuracy and dependable probability estimates. These qualities make it a strong candidate for future integration into intelligent clinical decision-support systems aimed at early detection and improved management of cardiac diseases.

8. FUTURE WORK

Although the proposed multimodal deep learning framework shows promising results for detecting Rheumatic Heart Disease, several areas still require further research and development. Broadening the scope of the study will help enhance the system's reliability, general applicability, and clinical value.

An important area for future research is multi-center validation. The dataset used in this study mainly comes from specific acquisition protocols and clinical conditions. Testing the proposed model across different healthcare institutions with various imaging equipment, patient populations, and clinical practices would strengthen evidence of its general use. Multi-center validation studies would also help identify potential biases and enhance the model's reliability in diverse clinical settings.

Another promising approach is integrating federated learning frameworks. Medical data often resides across many healthcare institutions, and privacy laws can restrict centralized data sharing. Federated learning enables models to be trained collaboratively on distributed datasets without transferring sensitive patient information to a central location. Incorporating federated learning into the proposed framework could facilitate large-scale model training while safeguarding patient privacy and complying with healthcare data laws.

Future work may involve deploying the proposed framework on portable or point-of-care ultrasound devices. Advancements in embedded hardware and edge computing enable the integration of deep learning models into lightweight medical tools. By optimizing the architecture for real-time operation and reducing computational demands, the system can be used in portable diagnostic devices to support cardiac screening in rural or resource-limited healthcare environments.

Additionally, incorporating advanced explainable artificial intelligence (XAI) methods is another important research area. While Grad-CAM offers useful visual explanations of model predictions, more sophisticated interpretability techniques could provide deeper insights into multimodal decision-making. Techniques such as attention visualization, concept-based explanations, and counterfactual analysis may help clinicians understand the reasoning behind automated predictions and foster greater trust in AI-supported diagnostic systems.

These future research directions aim to advance the current framework from trial phases to broader clinical implementation and real-world application. By incorporating large-scale validation, privacy-preserving learning methods, edge-device integration, and enhanced interpretability, the proposed system can become a dependable and scalable solution for automated cardiac disease screening.

REFERENCES

Rwebembera, J., et al. 2023 World Heart Federation guidelines for the echocardiographic diagnosis of rheumatic

heart disease. Nat. Rev. Cardiol. 2024, 21, 250-272.
Brown, K., Roshanitabrizi, P., et al. Using artificial intelligence to detect rheumatic heart disease via echocardiography: Focus on mitral regurgitation. J. Am. Heart Assoc. 2024, 13, e031257.
Milutinovic, S., et al. Rheumatic heart disease burden: A comparative analysis. JACC Adv. 2024, article 101393.
Zhang, J., et al. Recent advances in the prevention and management of rheumatic heart disease. Glob. Heart 2025.
Olatunji, D.E., et al. Machine learning-based analysis of ECG and PCG signals for detecting rheumatic heart disease: A scoping review (2015-2025). 2025.
Nakagaayi, D., et al. A curriculum to train trainers for scaling up artificial intelligence-assisted echocardiographic screening for rheumatic heart disease. 2025.
Qayyum, S.N., et al. A review of artificial intelligence applications in echocardiography. 2024.
Darwich, F., et al. AI-supported echocardiography for detecting heart disease: A review. 2024.
Hirata, Y., et al. AI in echocardiography: Current automated measurement techniques. 2025.
Sahashi, Y., et al. AI in echocardiography: Current status and future direction. 2025.
Myhre, P.L., et al. Artificial intelligence in echocardiography for cardiovascular medicine. 2025.
Maturi, B., et al. The role of artificial intelligence in echocardiography. 2025.
Kusunose, K., et al. The role of artificial intelligence in improving diagnostic performance in echocardiography. 2025.
Seetharam, K., et al. Expanding the view of artificial intelligence in echocardiography. 2024.
Shaulian, S.Y., et al. Integrating artificial intelligence into cardiac echocardiographic practice. 2025.
East, S.A., et al. Artificial intelligence-enabled point-of-care echocardiography and cardiovascular imaging. 2025.
Christensen, M., et al. Vision-language model for interpreting echocardiograms. Nat. Med. 2024.
Vukadinovic, M., Tang, X., Yuan, N., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D. EchoPrime: A multi-video view-informed vision-language model for interpreting echocardiography. arXiv 2024, arXiv:2410.09704.
Kim, S., et al. EchoFM: Foundation model for echocardiography video representation and analysis. arXiv 2024, arXiv:2410.23413.
Sahashi, Y., et al. Automating echocardiographic measurements with artificial intelligence. J. Am. Coll. Cardiol. 2025.
Hauptmann, T., et al. Echocardiographic readings by artificial intelligence compared to manual assessment. 2025.
Zhang, R., et al. Multimodal artificial intelligence in medicine: A task-based review. 2026.
Azarfar, G., et al. Responsible use of multimodal artificial intelligence in healthcare: Promises and challenges. Lancet Digit. Health 2025, 7, 100917.
Schouten, D., et al. Exploring the landscape of multimodal AI in medicine. 2025.
Li, Y., et al. A review of deep learning-based information fusion for medical classification tasks. 2024.
Zubair, M., et al. A detailed review of techniques, algorithms, advancements, challenges, and clinical uses of multimodal medical image fusion for better diagnosis. 2025.
Haq, I.U., et al. Advances in medical radiology through multimodal machine learning. 2025.
Huang, J., et al. The use of artificial intelligence in medical imaging for diagnosis: Recent advances and challenges. 2025.
Partovi, E., et al. A review of deep learning methods for analyzing heart sound signals. 2024.
Orozco-Reyes, L. et al. A deep-learning method for classifying heart sounds using time-frequency representations. Technologies 2025, 13, 147.
Liu, N., et al. A deep learning model for heart sound classification using dual-branch fusion. 2026.
Althaph, B., Challa, N.P. Explainable attention-based deep learning for classifying and interpreting heart murmurs using phonocardiograms. Sci. Rep. 2025, 15, 37991.
Li, M., et al. Heart sound classification using multi-scale feature extraction and channel attention fusion. 2025.
Muhammad, D., et al. Understanding the black box: A systematic review of explainable artificial intelligence in medical image analysis. 2024.
Houssein, E.H., et al. Explainable artificial intelligence for medical imaging with deep learning: A comprehensive review. Cluster Comput. 2025.