RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2622

HTCTA: A Hybrid Transformer Based Temporal Attention Mechanism for Early Diagnosis of Cardiovascular Abnormalities from PCG Signals

Bollapalli Althaph [a], Nagendra Panini Challa [a,*]

[a] School of Computer Science and Engineering (SCOPE), VIT-AP University, Beside AP Secretariat, Amaravati, 522241, Andhra Pradesh, India

* Corresponding Author Email: nagendra.challa@vitap.ac.in

DOI: https://doi.org/10.54392/irjmt2622

Received: 28-08-2025; Revised: 22-01-2026; Accepted: 09-02-2026; Published: 02-03-2026

Abstract: Cardiovascular diseases are the leading cause of death across the world and responsible for about one third of all deaths. It is important to detect heart problems early and accurately, before serious damage can occur. Recording of the heart sounds, known as phonocardiograms (PCGs), is a non-invasive and inexpensive method for diagnosis. Nevertheless, the natural non-stationarity, noise and variability of PCG signals remain as a grand challenge for traditional DL methods. Although convolutional neural networks can effectively model the local features of spectrograms, it is difficult to model long-term dependencies in spectral feature maps. On the one hand, transformer based models may model temporal relationships but likely do not possess the capability to localize fine grain clinical patterns. To address the above problems, in this paper we introduce a Hybrid Architecture of Transformer-CNN with Temporal Attention (HTCTA). The model consists of CNNs in capturing localized time-frequency features, temporal attention to highlight diagnostically significant cardiac segments (e.g., systole or diastole), and Transformer encodings in pooling long-range dependencies over the heart cycle. Mel-spectrograms processed from heart sound recordings are forwarded through the hybrid model for classification. The proposed HTCTA model was tested over the two benchmark datasets, namely, PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022. It reached a classification accuracy, precision, recall and F1-score of 94.70%, 94.20%, 95.15% and 94.67%, respectively, outperformed a number of state-of-the-art models, including Whisper-based as well as CRNN architectures. The model is moreover resistant to noise and variability between (auscultation) positions. Because of negligible difference between the reference and restored data, by virtue of being accurate, interpretable, efficient, HTCTA has potential in its real-time clinical diagnosis and medicine application. In the future, multimodal inputs will be incorporated and cross-patients validation will be performed to improve the generalization.

Keywords: Cardiovascular Disease (CVD), Heart Sound Classification, Phonocardiogram (PCG), Temporal Attention, Transformer Neural Network, Convolutional Neural Networks (CNNs), Hybrid Deep Learning.

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, responsible for over 17.9 million deaths annually and accounting for nearly one-third of all global fatalities [1]. Early detection of pathological cardiac conditions plays a critical role in reducing morbidity and improving clinical outcomes. Among different non-invasive diagnostic techniques, phonocardiogram (PCG) is an affordable, portable, and radiation-free approach for the detection of heart function [2]. PCGs and digital stethoscopes can provide widespread access for screening in primary-care, rural, and mobile health settings where echocardiographic services are not readily available [3]. Nonetheless, manual auscultation is subjective and subject to bias between observers, underpinning the need for automated machine learning- or deep learning-based systems that could aid clinical decision-making [4].

Nevertheless, precise automatic analysis of PCGs is still a difficult task, because the heart sound signals are inherent with non-stationarity and overlapping frequency components, background noise and also due to the heterogeneity in recording devices and auscultation positions [5]. Very recent work in this area has focused on the impact of heterogeneity: different sensor response, noise conditions and sampling protocol can lead to large decrease in model generalization capability [6]. For the latter, considerable

enhancing on balanced accuracy was achieved by using DMD and heterogeneity-resilient training approaches [7]. In addition, the deep ONMF [8] has demonstrated its strong discriminative power in PCG spectrogram analysis.

Conventional ML algorithms that rely on MFCC, wavelet transforms, scattering transforms, and logistic regression have been used as benchmarks for murmur classification [9]. For instance, wavelet scattering transforms together with 1D-CNNs have achieved state-of-the-art performances on benchmark datasets [10]. However, handcrafted feature dependence restricts the adaptability, and traditional classifiers are unable to capture the long-range dependencies and temporal changes of PCG signals.

Networks such as CNN and RNN, achieved good performances in the murmurs, valvular disorders and pathological heart sounds detection [11]. On one hand, CNN-based methods are good at modeling local spectro-temporal information in Mel-spectrograms yet limited by fixed receptive fields. RNNs and Bi-LSTMs take into account the sequential dependence but generally suffer from vanishing gradients, increased computation overhead and less deployability [12]. Hybrid CNN-LSTM architectures and attention mech- anism have also been used to enhance sequential modeling and rank the state-of-the-art performance in PhysioNet and CirCor data- sets [13]. Additionally, multimodal fusion methods using ECG and PCG have reported an accuracy of more than 97%, showing that it is beneficial to combine the electrical and acoustic biomarkers for prediction [14].

Transformer models has become popular for the purpose of temporal attention and modeling long-range dependencies in recent years. Recently proposed Multi-head Attention, Pyramid-Dilated CNNs and deep frequency modeling outperformed frame-wise detection performance of murmurs on benchmark data sets [15]. However, without localized feature extraction, pure transformer architectures can ignore fine-grain spectral clues. On the other side, CNN models cannot capture intra-cardiac level dependencies without using recurrent or attention-based enhancement. Therefore, a significant research gap remains for models that are able to aggregate local, global and clinically meaningful features into a single framework.

In recent years, the transformer models have been widely used for temporal attention and to model long-term relationship. The state-of-the-art algorithms recently proposed in Multi-head Attention, Pyramid-Dilated CNNs and deep frequency modeling surpassed the frame-wise detection performance of murmurs on benchmarked datasets [16-18]. However, the pure transformer can be difficult to handle fine-grain spectral clues without performing local feature extraction. On the flip side, CNN models are unable to learn within intra-cardiac level dependency without explicitly using RNN or attention-based enhancement. Thus there is a clear research challenge for models that can encode local, global, and clinically relevant information in the same framework.

To overcome these drawbacks, we present a Hybrid Transformer-CNN with Temporal Attention (HTCTA) model. HTCTA combines (i) convolutional layers for localized time–frequency feature extraction, (ii) a temporal attention mechanism to highlight diagnostically informative cardiac segments and (iii) transformer encoders to capture long-range dependencies across heart cycles. The model seeks to address the restrictions of both CNN-only and transformer-only, by integrating spatial and global features together into a unified light-weight pipeline. In contrast to the multimodal ones, HTCTA only uses PCG signals and thus can easily coupled with digital stethoscope and health embedded device.

The proposed model serves four aspects of the current needs:

• Generalization between heterogeneous datasets (Physio Net and CirCor),

• Temporal interpretability through attention,

• Computational speed appropriate for real-time inference, and

• Insensitivity to noise and variation in auscultation.

Literature review Section 2 presents the related work which explores the limitations of existing techniques. In Section 3, we present framework and major elements of it. Experimental results along with analysis are presented in Sec. 4. Section 5 and Section 6 contain main findings, limitations, contributions and future work.

2. Literature Review

In recent years there have been a variety of computational techniques proposed to enhance the diagnostic potential for CVD diagnosis through analysis of HS and ECG signals. The handbook reviews what can be considered a progressive development of the literature from traditional classification approaches toward more sophisticated computational methodologies which make use of optimal feature extraction, noise-tolerant processing and improved decision making schemes. Cumulatively, the study aims to deal with some practical challenges including variability in signal acquisition, data imbalance and limited hardware resources as well as effective real time monitoring in clinic.

To compare these developments systematically, Table 1 Comparative Literature Review on Heart Disease Classification Techniques is presented to outline the major methods and optimization techniques used, as well as the results achieved by the studied

contributions. This integrated perspective underscores ongoing developments, identifies existing research challenges and the growing relevance of intelligent diagnostic systems to facilitate early detection, risk assessment and improved decision support in cardiovascular care.

Table 1. Comparative Literature Review on Heart Disease Classification

S.No.	Authors/ Year	Title	Technique	Key Contribution	Accuracy
1	T. Priyadarsini, P. K. Sahu, S. M. Dey U. / 2024 [19]	A Novel Optimized Machine Learning Approach for Early Prediction of Heart Disease Using Bio-Inspired Algorithms	Genetic Algorithm, Bat Algorithm, Ant Colony Optimization	Enhanced ML accuracy through optimized feature selection.	94.56%
2	Rui Yi, Li Wang, Zhihao Cao, Jian Huang / 2024 [20]	Optimization of Transformer Heart Disease Prediction Model Based on Particle Swarm Optimization Algorithm	Particle Swarm Optimization (PSO)	PSO-optimized Transformer outperformed Random Forest.	92.01%
3	Saeid Veisi, Saeed Khoshhal, Saeed Jalili / 2023 [21]	Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm	Jellyfish Optimization Algorithm	Improved MLP accuracy via feature selection.	98.47%
4	Saravana Balaji B, Ashokkumar P, Saifur Rahman / 2023 [22]	Intelligent Bi-LSTM with Architecture Optimization for Heart Disease Prediction in WBAN	Improved Dingo Optimizer (IDOX)	Enhanced Bi-LSTM accuracy using optimized architecture.	89.21%
5	Mohammad Rafidul Islam, Iftekharul Mobin, Nabeel Al-Qirim / 2025 [23]	CardioTabNet: A Novel Hybrid Transformer Model for Heart Disease Prediction Using Tabular Medical Data	Random Forest Feature Ranking + TabTransformer	Achieved 94.1% accuracy with hybrid deep model.	94.1%
6	Yogesh Aher, S. S. Thakur, K. M. Bhurchandi / 2025 [24]	Squeeze RNN with Hybrid Optimization for Heart Disease Prediction Using Gene Expression Data	Hybrid Optimization	Accurate prediction using optimized Squeeze RNN.	93.20%
7	P. B. Reddy, N. Subashini, S. K. Patel / 2022 [25]	An Efficient Prediction System for Coronary Heart Disease Risk Using Selected Principal Components and Hyperparameter Optimization.	PCA + Hyperparameter Optimization	Achieved strong results through dimensionality reduction and tuning.	94.31%
8	Abdul Jabbar *et al.* / 2025[1]	Automated Detection of Pediatric Congenital Heart Disease from Phonocardiograms using Deep and Handcrafted Feature Fusion	Deep feature fusion (handcrafted + DL)	Early CHD detection using fusion of engineered + learned features for improved generalization.	92%

9	Ehsanhosein Kalatehjari *et al*. / 2025 [2]	Advanced Ensemble Learning-based CNN-BiLSTM Network for Cardiovascular Disease Classification Using ECG and PCG	Ensemble learning	Combines ECG and PCG in hybrid ensemble for improved CAD diagnosis.	97%
10	Adrian Florea *et al*./ 2025 [3]	Exploring Finetuned Audio-LLM on Heart Murmur Features	Finetuned Audio-LLM	Uses Qwen2-Audio LLM to classify murmur features with enhanced noise robustness.	Not specified (outperforms SOTA)
11	Nadikatla Chandrasekhar *et al*. / 2025 [4]	Heart Abnormality Classification Using ECG and PCG Recordings with Novel PJM-DJRNN	Poisson-based feature selection + PDF-SLO	Novel recurrent model using dual-signal fusion improving heart abnormality classification.	97.33%
12	Monjur Morshed & S.A. Fattah *et al* . / 2022 [26]	Deep Learning-based Murmur Detection using Joint Optimization and Decision Fusion	Decision fusion	Parallel networks for four valve-specific PCG improving murmur detection.	92.21%
13	V. Madhava Shervegar *et al* / 2025 [6]	Heart Sound Classification for Early CVD Detection using Improved WST & DA Classifier	Wavelet scattering + discriminant analysis	Combines 2D-DCT + improved WST for lightweight CVD classification.	99.63%
14	Ahmed Patwa *et al.* / 2025 [13]	Heart Murmur & Abnormal PCG Detection via WST and 1D-CNN	Wavelet scattering transform + 1D CNN	Preprocessing + WST improves patient-wise murmur classification.	Weighted accuracy competitive; experiment-specific (up to SOTA)
15	E. A. Nehary & S. Rajan *et al*. / 2025 [14]	Phonocardiogram Classification Using DMD for Heterogeneity-Resilient Training	Dynamic Mode Decomposition + domain-balanced training	Addresses sensor heterogeneity to improvebalanced accuracy across datasets.	~15% improvement in balanced accuracy
16	Shubham Basak & Ujjal Bhattacharya *et al* / 2025 [15]	Deep Determination of Cardiac Condition from Phonocardiograms	Pyramid dilated convolution + multi-headed attention	Enhances CAD and murmur recognition using multi-scale learning.	Improved over SOTA (not exact %)
17	Samira Moghani *et al.* / 2025 [27]	Enhanced Heart Sound Analysis using Hierarchical Spectral Basis via Deep ONMF	Deep Orthogonal Non-Negative Matrix Factorization	Robust PCG feature extraction with statistical separability even in noise.	CI 95.2–97.8%
18	Zhang, H *et al*. / 2025 [28]	Valvular Heart Disease Classification	DWT + WPT + NCA + SVM/RF/KNN	PCG-based AS/MS diagnosis	SVM: 97%, RF: 96.2%
19	M. Morshed & S.A. Fattah *et al*. / 2025 [5]	Four-Valve Murmur Detection	Parallel DL networks + decision fusion	Multi-lead PCG improves murmur detection	92.21%

20	Adrian Florea *et al.*/ [3]	Audio-LLM for PCG Feature Classification	Fine-tuned Qwen2-Audio + SSAMBA segmentation	LLM achieves SOTA in 10/11 tasks	Outperforms baselines (no explicit %)
21	Vijayasimha, A *et al*. / 2025 [29]	Hybrid deep learning framework for cardiovascular disease diagnosis and prognosis using GAN, LSTM, GRU, VARMA, and deep DynaQ network	CNN + RNN + RL Fusion	Multimodal PCG+ECG classification	95%
22	Talal, M *et al*. / 2023 [30]	Machine learning-based classification of multiple heart disorders from PCG signals	SVM, Fine-KNN	Identifies optimal ML-feature pairs	98.3%
23	Kriti Taneja *et al.*/ 2023 [31]	Classifying the heart sound signals using textural-based features for an efficient decision support system	HOG + LBP + kNN	binary pattern (LBP), adaptive-LBP, and ring-LBP	95.27 %
24	Ahmed Patwa *et al.*/ 2025 [13]	Heart murmur and abnormal PCG detection via wavelet scattering transform & a 1D-CNN	WST + 1D-CNN	Noise-robust murmur detection	Weighted accuracy SOTA (no exact %)
25	Ebrahim Nehary & Sreeraman Rajan *et al.* / 2025 [14]	Phonocardiogram classification using dynamic mode decomposition for heterogeneity-resilient training	Dynamic Mode Decomposition + domain-balanced training	Handles sensor variability	~15% improvement in balanced accuracy
26	Shubham Basak & Ujjwal Bhattacharya *et al*. (2023 [15]	Pyramid Multi-branch Fusion DCNN with Multi-Head Self-Attention for Mandarin Speech Recognition	PDA + multi-head attention + focal-triplet loss	CAD + murmur detection outperforming SOTA	Significant improvement (exact % not stated)
27	Samira Moghani *et al.* / 2025 [27]	Enhanced heart sound analysis through hierarchical spectral basis vector extraction using deep orthogonal non-negative matrix factorization	CNN	Discrete Wavelet Transform (DWT), Mel-Frequency Cepstral Coefficients (MFCC),	95.2–97.8%

3. Dataset and Distribution

In this study, the proposed HTCTA model was tested on two publicly accessible benchmark datasets: (i) the PhysioNet/Computing in Cardiology Challenge 2016 database and (ii) the CirCor DigiScope 2022 data14. Both datasets include heart sound recordings obtained at different auscultation sites (i.e., aorta, pulmonic, tricuspid, and mitral) and have natural variations in signal quality, noise characteristics, and on patient demographics. In order to avoid being biased toward a particular recording protocol or cohort, all experimentation were conducted with patient-level data splits so that there was no overlap between the training and test sets.

Class distribution on both datasets is naturally unbalanced, as murmurs are underrepresented relatively to normal sounds. To overcome this issue, stratified sampling was used to keep normal and abnormal case ratios in each fold. Furthermore, oversampling for minority classes was used when

appropriate to avoid a biased model towards the majority class during training. The dataset was divided into training, validation and test sets using a 70%–15%–15% split ratio in which each patient's recordings were not be shared across the sets.

To validate generalization and address dataset-specific bias, we applied a 5-fold stratified cross-validation approach. The model was trained on four folds and the fifth one was used for testing, average performance in terms of both precision and recall is reported over all cross-validation sets. Such cross-validation is especially important for PCG databases, as inter-patient variability and different recording conditions are present.

The HTCTA was trained using an Adam optimizer with the learning rate of 0.001 and a cosine decay learning rate schedule for convergence in the long run. A mini-batch size of 32 was used, and early-stopping criteria was included in the training process with a maximum of 100 epochs to prevent overfitting. Dropout regularization and batch normalization were included into the training pipeline for further generalization improvement. All our experiments were run on a GPU-based setup and the computation time was measured for realtime deployement. Thus, this section aims at making the experimental setup fully reproducible and transparent by explicitly stating dataset balancing, patient-level splitting, cross-validation method, and training hyperparameters.

4. Methodology

The proposed HTCTA model is trained and validated on two benchmark datasets: PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022 dataset. These are made up with PCG datasets which were acquired on different populations and locations for auscultation. PCG files in. wav files with labeling (e.g., normal, abnormal). Librosa or Scipy load in these data sets. io. wavfile which we can use to read raw audio as wave arrays and associated sample rate. Each audio file is matched with its corresponding label, forming a structured dataset. A pipeline is introduced for iterating through folders, signals extracting, labeling assigning and saving of preprocessed segments in memory efficient forms (HDF5 files or NumPy arrays). The resulting segments are then converted to Mel-Spectrograms (a time-frequency representation that reflects the human auditory system). The Short-Time Fourier Transform of a signal a (t) is initially computed:

$$A(n,c) = \sum_{n=-\infty}^{\infty} a(n)c(n-c)e^{-jcn} \qquad (1)$$

Where c (n) is a window function centred at time m. The Mel-spectrogram N (g, v) is then computed using a Mel-scale filter bank applied to

$$\mathrm{N(g,v)} = \mathrm{MelFilterBank.}\,|\mathrm{A(v,g)}|^2 \qquad (2)$$

After the audio tracks are loaded, they are preprocessed in order to increase their quality and consistency. The signals are bandpass filtered (25–400Hz) to capture the frequency components of S1, S2, murmurs and other heart sound related sounds garbage that removes low frequency drift and high frequency environmental noise. Afterward, the signals are centered and scaled to zero mean and unit variance. Due to variation in heart sound lengths, recordings are divided into 4–6 second segments using the sliding window method with overlapping (e.g., 50%). The concatenation is transformed into a Mel-spectrogram, which represents energy over time and frequency in terms of the human perception (Mel-scale). The resulting spectrograms are resized (e.g. 128×128) and normalized to lie in the range 0–1 always for the model input.

The feature map of improved l_{th} layer convoluting is:

$$G^{l} = \sigma(\mathrm{DN}(\mathrm{C}^{(l)} * \mathrm{G}^{l-1} + \mathrm{b}^{(l)})) \qquad (3)$$

Where * stands for convolution, for ReLU activation, $C^{(l)}$, and $b^{(l)}$ are weights and bias separately, and DN is batch normalization. In order to design a successful automated decision-making system capable of identifying normal and abnormal heart sounds (HSs), one needs step-by-step approach that includes preprocessing, feature extraction, segmentation and classification. This way it guarantees that raw PCG signals are transformed into exploitable and consistent features that can be effectively fed to machine learning algorithms. The robustness and classification capability are dependent on the quality of signals, recorders used, patient state or conditions; thus making preprocessing and design of accurate feature essential. This section describes each step in the processing pipeline, beginning with audio preprocessing.

Figure 1 shows the end-to-end pipeline we followed, which processes raw heart sound recordings to detect beats and transform them into Mel-spectrograms. We propose to design parallel CNN and Transformer branches that learn to complement spectral and temporal representations, which can be effectively fused and refined with a temporal attention mechanism for normal/abnormal classification.

Audio Filtering: Raw waveform PCG Generally, raw waveforms PCGs are preprocessed through several procedures to make consistent quality and facilitate feature extraction. First, each signal is loaded and, if necessary, resampled to a common rate. Down sampled all recordings to 1000 Hz using an anti-aliasing filter. Keeping a standard sampling rate (e.g. 1000 Hz) simplifies later processing. Signals are then normalized so that amplitude differences between recordings do not bias the features.

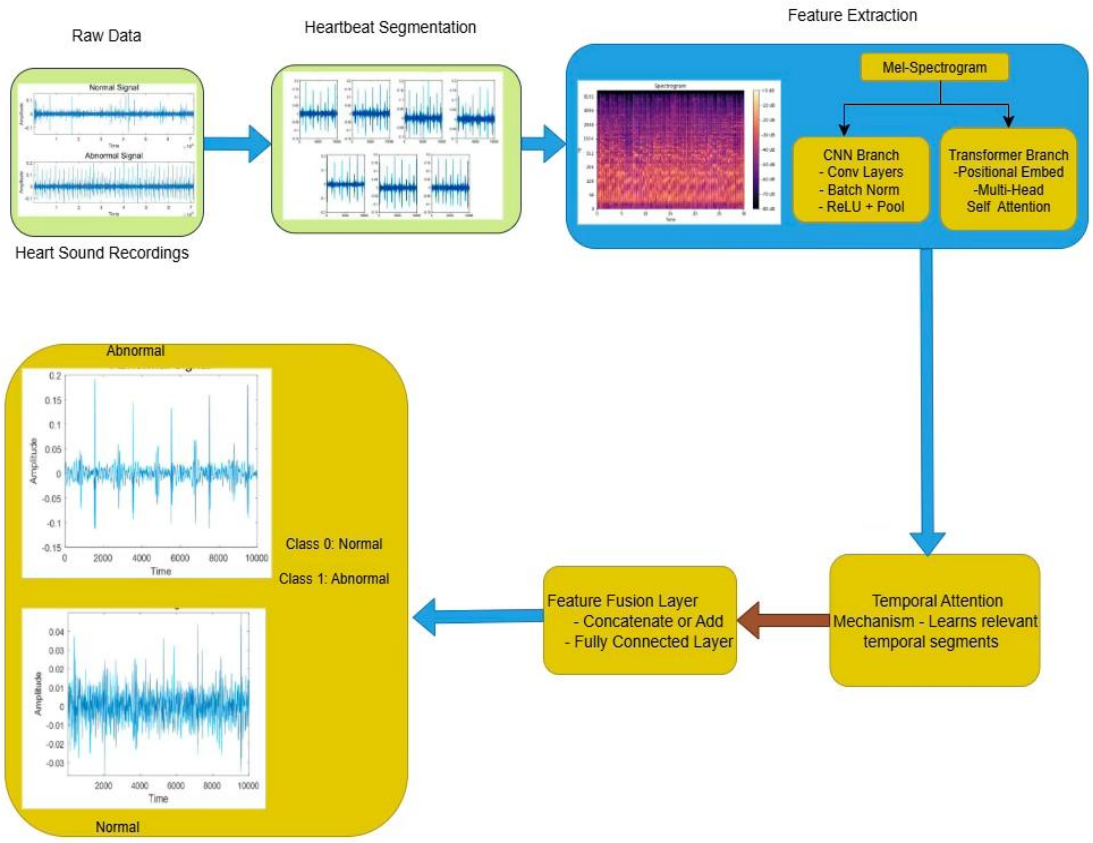

Figure 1. Methodology for Heart Disease Prediction Using Pipeline

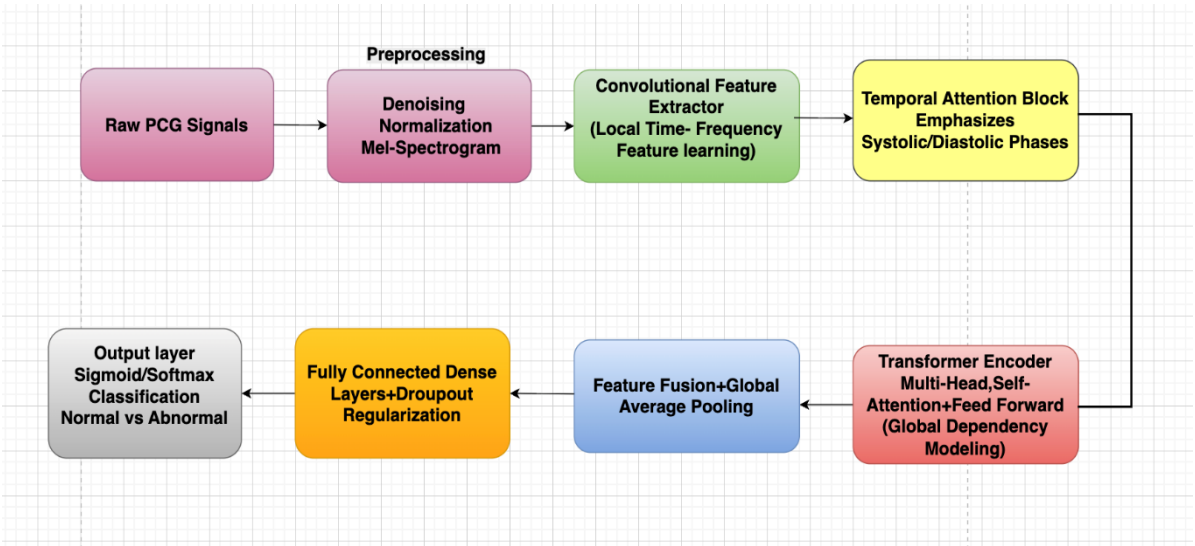

Figure 2. Proposed HTCTA Architecture for PCG Classification

Let the CNN feature map be F= [f1, f2, f3…..f_t] $\in S^{cXf}$, where T is the number of time steps, and d is the feature dimension. The attention score for time step t is computed as:

$$\alpha_t = \frac{\exp(e_v)}{\sum_{i=1}^{V} \exp(e_i)} \quad (4)$$

The weighted feature representation is:

$$\hat{F} = \sum_{v=1}^{V} \alpha_v f_v \quad (5)$$

Next, a bandpass filter is applied to remove irrelevant frequencies. Heart sounds predominantly lie in roughly the 20–200 Hz range; higher-frequency components (like speech) and very low-frequency trends are considered noise. A typical approach is a Butterworth bandpass (e.g. ~20–200 Hz) which greatly attenuates voice and muscle noise. Figure 2 compares

a raw PCG waveform (black) to the same signal after bandpass filtering (red). The filter clearly suppresses much of the low-frequency baseline wander and high-frequency spikes, isolating the heart sound oscillations [5].

Figure 2 illustrates the overall architecture of the proposed Hybrid Transformer-CNN with Temporal Attention (HTCTA) model. The system starts by obtaining raw PCG signals in. wav format, which are postprocessed by denoising, normalization, and Mel-spectrogram conversion afterwards. This preprocessing serves to reduce noise artifacts and captures the acoustic structure of heart sounds for later learning. The output spectrograms are then fed into the convolutional feature extractor, in which ConV2D, Batch Normalization, ReLU activation function, and Max-Pooling layers collaboratively learn localized time–frequency patterns related to critical cardiac events such as S1 and S2 sounds or murmur features.

After CNN stage, the generated feature maps are passed into a Temporal Attention Block (TAB) which will explicitly focus on diagnostically relevant systolic and diastolic phases, to allow the network to emphasize clinical important intervals more than background ones. The attention-weighted representations are further fed into a Transformer Encoder which takes advantage of multi-head self-attention and feed-forward modules to handle long-range dependencies among cardiac cycles. This layer allows the network to perceive global temporal patterns that could reveal pathological anomalies.

The features are pooled with a Feature Fusion and Global Average Pooling (GA) module to reduce dimension and retain useful contextual information after GCL. The concatenated representation passes through dense layers with dropout regularization to avoid overfitting. Lastly, the output layer uses Sigmoid or Softmax activation based on binary or multi-class formulation for classification of heart sound signals into Normal and Abnormal classes. In summary, we combine localized CNN learning, temporal attention gating and transformer-based global modeling to perform robust and interpretable PCG classification.

From Figure 3 after filtering, signals are often segmented into cardiac cycles or windows for analysis. One goal may be to isolate individual heartbeats or the first/second heart sound events (S1, S2) as separate segments (see Segmentation section). Even without explicit segmentation, signals are typically split into fixed-size frames (e.g. 25 ms with overlap) during feature extraction. Any remaining silence or extremely low-energy portions can be trimmed or ignored. Finally, one may apply additional normalization (for example, z-scoring or equalizing each frame) to reduce session-to-session variability. Preprocessing ensures that each heart sound recording is at a consistent sampling rate, has a standardized amplitude range, and has out-of-band noise removed. These steps greatly improve the robustness of feature extraction and downstream classifiers.

Multiple heads allow the model to jointly attend to information from changed subspaces:

$$\text{MultiHead}(A) = \text{Concat}(\text{head1},...,\text{headh})A^{O} \quad (6)$$

Feature Extraction: Once the raw PCG is pre-processed, the next step is to compute features that capture the salient characteristics of heart sounds. The most common features for audio signals are Mel-Frequency Cepstral Coefficients. MFCC extraction proceeds as follows: (1) apply a pre-emphasis filter to balance the spectrum; (2) split the signal into short, overlapping frames (e.g. 25 ms with 10 ms step) and apply a window (e.g. Hamming) to each frame; (3) compute the power spectrum of each frame via an FFT; (4) pass the power spectrum through a set of mel-scale triangular filters; (5) take the logarithm of the filter outputs; and (6) apply a Discrete Cosine Transform (DCT) to de-correlate and compress them, keeping typically the first 12–13 coefficients. This yields a sequence of MFCC vectors over time. The mel filter bank step is designed to mimic human hearing sensitivity: filters are narrowly spaced at low frequencies and more widely spaced at high frequencies.

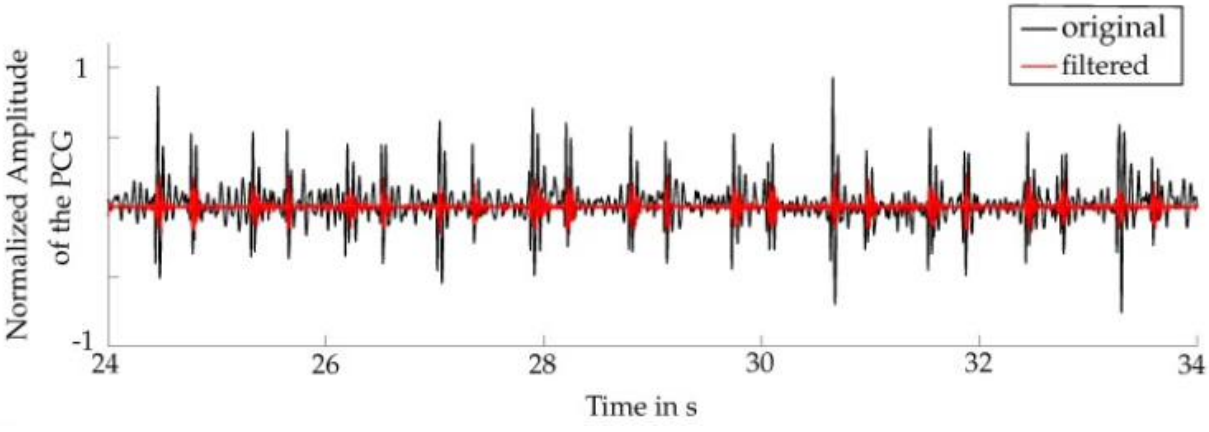

Figure 3. Waveform before (black) and after (red) bandpass filtering [12]

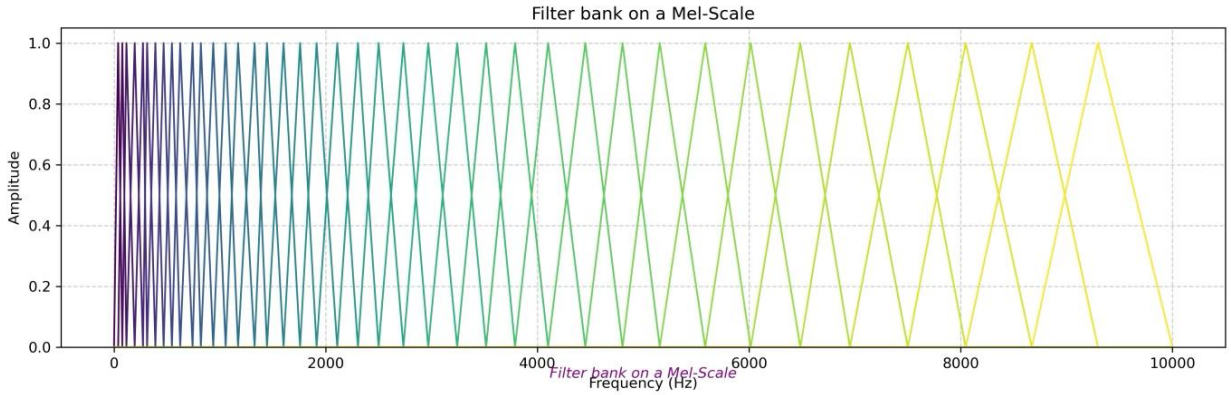

Figure 4. Illustration of a bank of triangular mel-scale filters spanning 0–10000 Hz. Each filter's output (on overlapping bands) is computed from the frame's FFT magnitude [16]

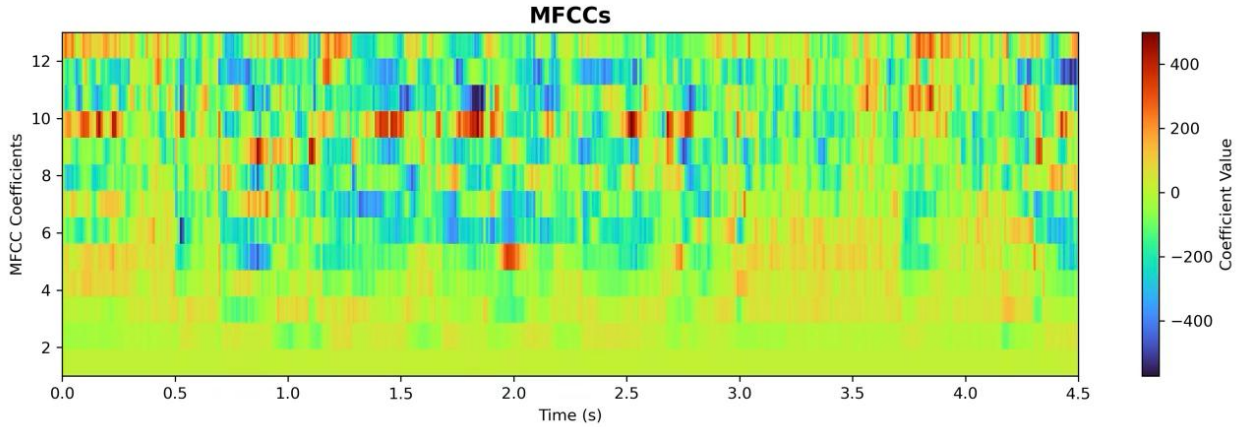

Figure 5. Example MFCC spectrogram (12 coefficients over time) for a 3-second heart sound segment. (Warm colours = higher coefficient values.)[15]

Figure 4 shows an example mel filter bank. After applying these filters and taking log, a DCT projects onto a set of orthogonal "cepstral" bases; the first 12–13 DCT coefficients per frame are taken as MFCC features. The result is a two-dimensional feature matrix (time vs. cepstral index). Figure 4 shows a typical MFCC heatmap extracted from a heart sound.

From Figure 5 MFCCs capture the overall spectral envelope and are robust to small shifts in frequency content. They are by far the most popular acoustic features in heart sound analysis. In addition to MFCCs, several supplementary features can enhance classification. Common examples include:

- Spectral features: e.g. spectral centroid (the "centre of mass" of the spectrum), bandwidth, spectral roll-off, and chroma (energy in musical pitch classes). These describe how frequency content is distributed and have been used in audio analysis more broadly.
- Time-domain features: Zero-Crossing Rate and Root-Mean-Square energy. ZCR measures how often the signal changes sign per frame; a higher ZCR often indicates noisy or pathological sounds. RMS energy is proportional to the frame's power; abnormal PCGs (with murmurs) typically have higher overall power than normal ones [7].
- Temporal features: e.g. the durations of S1, S2, systole, and diastole. Longer-than-normal durations of certain states can be indicative of pathology. These are obtained after segmentation (see next section).
- Entropy or waveform shape features: e.g. Shannon entropy or kurtosis of the waveform, which reflect signal complexity and have been explored in some studies moody-challenge.physionet.org

Each feature vector (MFCCs plus any additional features) is then paired with the recording's label (Normal/Abnormal) to form the dataset for classification.

Let $E \in S^{cXf}$ be the transformer output, then:

$$T = \frac{1}{V}\sum_{v=1}^{V} Z_v \quad (7)$$

$$\hat{y} = \mathrm{Softmax}(C_t z + b_t) \quad (8)$$

Depending on whether the task is binary classification. Feature Engineering and Dataset Preparation: After extraction, all features are assembled into a structured dataset. Each PCG recording (or window/segment of it) corresponds to one feature vector (MFCC coefficients, spectral and temporal features) and one label. Typically, categorical labels (Normal/Abnormal) are encoded numerically (e.g. 0/1) or in one-hot form for use in classifiers. The feature matrix is often stored in a convenient format (e.g. CSV, NumPy array, or HDF5) with consistent indexing. Before training, the data is split into training and testing sets.

Algorithm 1: HTCTA Training Pipeline

1. Input: Raw PCG signal a(c), labels b
2. Pre-process: Filter → Normalize → Mel-spectrogram N(g(b))
3. CNN: Extract local features E∈S^{cXg}
4. Attention: Compute temporal weights α_c, obtain
5. Transformer: Learn global context Y= Transformer
6. GAP & FC: x=GAP(x) = , = Classifier (x)
7. Loss: Compute D(b,)
8. Back-propagation: Update weights using Adam optimizer
9. Repeat: Until convergence

A stratified split is recommended to preserve the Normal/Abnormal ratio in each subset, given the class imbalance. Many studies use K-fold cross-validation with stratification to reliably estimate performance. For example, a stratified 10-fold CV ensures each fold has ~79% Normal and 21% Abnormal data, matching the overall distribution. At each training fold, one can further set aside a validation split or use nested CV to tune model hyperparameters. When using complex preprocessing or feature pipelines (e.g. scaling, filtering, MFCC computation), it is good practice to serialize the pipeline. For instance, in scikit-learn one might create a Pipeline that applies the same normalization and transforms to any new data, and save it with joblib. This ensures that future data is processed identically to the training data. Likewise, label encoders or oversampling models (see next) should be preserved so that deployed systems remain consistent with the training phase. Once features and labels are prepared and split, the dataset is ready for modeling. Any further adjustments – such as oversampling minor classes or generating synthetic data – can be applied to the training set only, avoiding leakage to test data.

Segmentation Techniques: Heart sound signals have a well-defined structure: each beat contains two main sounds (S1 and S2) separated by systole and diastole intervals (Figure 5). Identifying the S1/S2 events or segmenting into individual cardiac cycles can improve analysis (e.g. by computing features for each cycle). Two common segmentation approaches are:

Hidden semi-Markov models (HSMM): Probabilistic models like Springer's HSMM have been widely adopted. These models treat the sequence of heart sounds as hidden states (S1, systole, S2, diastole) and use the signal's envelope as input. Given the trained state durations and emission probabilities, the HSMM algorithm finds the most likely state sequence. Others have used an improved version of Schmidt's HSMM method to segment each signal into the four states. This yields time indices for each S1 and S2 onset.

Envelope-based methods (Hilbert transform): Another approach is to compute the signal envelope (using the Hilbert transform or low-pass filtering the magnitude) and then detect peaks corresponding to heart sound events. For example, computing the absolute value of the analytic signal (Hilbert envelope) produces a smooth curve whose peaks align with S1/S2. Peak detection on this envelope (e.g. gradient zero-crossing) can then identify each sound. Comparative studies found the Hilbert-envelope method to be computationally efficient and nearly as accurate as more complex models.

In practice, HSMM tends to yield slightly higher segmentation accuracy in noisy conditions, but envelope-based methods are faster and simpler. Often, HSMM or Hilbert envelope algorithms are used during preprocessing to label state boundaries. Those indices can then be used to compute state-duration features or to align features across beats. Accurate segmentation is especially important if one wants to focus on S1/S2 sounds or detect abnormal sounds like murmurs that occur between S1 and S2.

Data Augmentation and Class Balancing: Given the limited and imbalanced nature of heart sound data, data augmentation and oversampling techniques are frequently used to improve classifier robustness. One common approach to address class imbalance is SMOTE, which creates imitation examples of the minority class by incorporating feature vectors. Beyond SMOTE on features, one can augment audio signals directly. Common augmentations in audio include:

Figure 6 illustrates the user-system interaction, that is, from detection of the onset of heart disease to data set retrieval from PhysioNet, as described above. It also provides a simplified pipeline for heart sound analysis which includes data preparation, audio preprocessing and MFCC feature extraction as the three essential steps.

Figure 7 shows a phonocardiogram waveform with the detected S1 and S2 instants, which are valve closure times. Systolic and diastolic duration intervals of S1–S2 and S2–S1 are denoted as a full heart cycle for heart beat segmentation and feature generation.

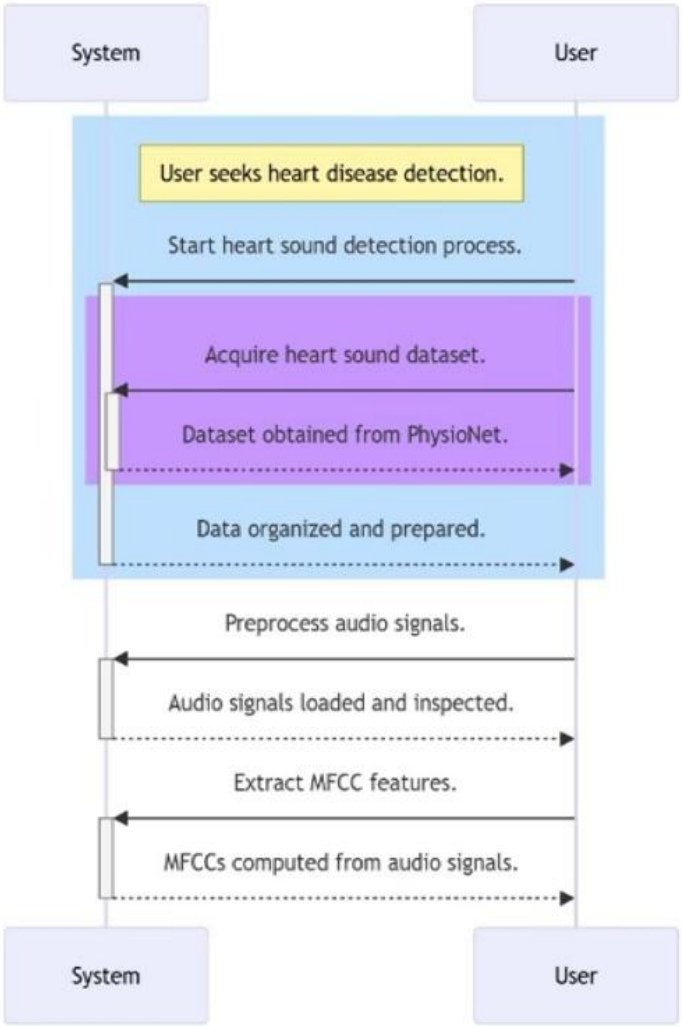

Figure 6. Heart Sound-Based Disease Detection Workflow

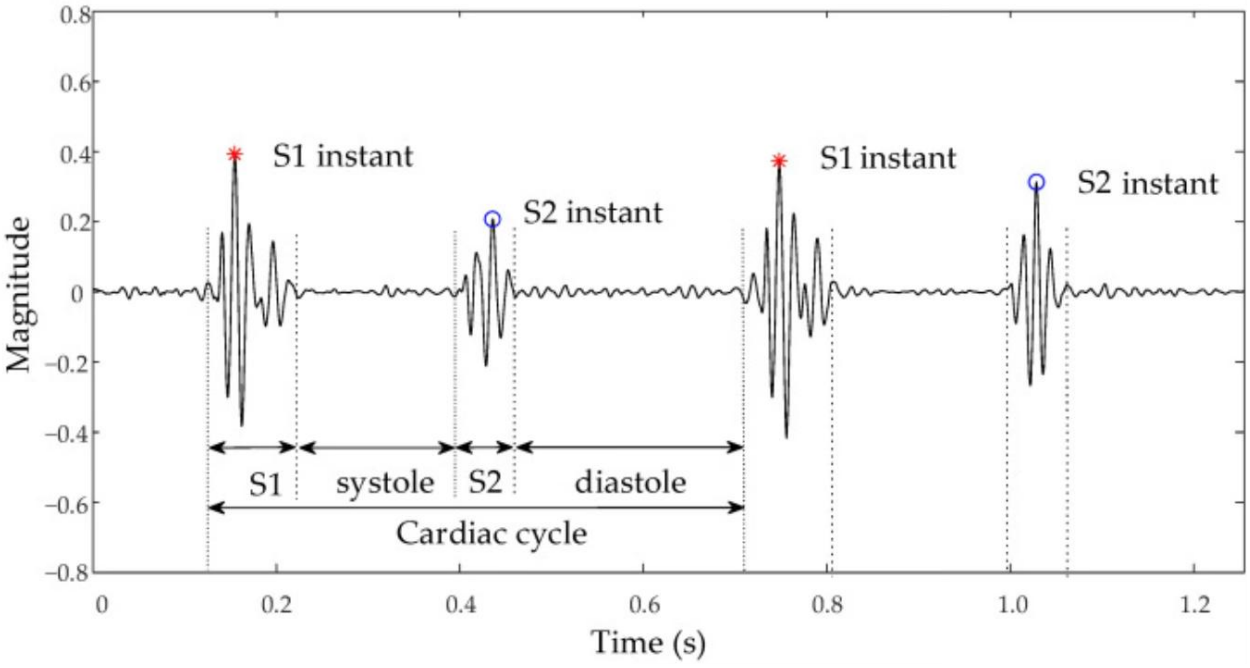

Figure 7. To illustrate the main heart sounds (S1, S2) and phases (systole, diastole) in a cardiac cycle, showing how these intervals can be used for cycle-based feature analysis in heart sound studies [15].

Pitch shifting: Changing the playback speed slightly or resampling to raise/lower pitch by a few semitones. This simulates variation in heart rate without altering timing too much. Time stretching (speed perturbation): Slightly compressing or stretching the audio in time (e.g. ±5–10% speed) simulates faster or slower heart rates (tachycardia or bradycardia). Noise addition: Adding a small amount of background noise (e.g. white noise or low-level random noise) can improve robustness. However, one must be cautious: unlike other audio domains, adding certain noises to heart sounds can mimic murmurs or distort the signal content. The augmentation should remain physiologically plausible. Window slicing / cropping: Randomly selecting different contiguous 5–10 second segments from longer recordings effectively increases sample count. When applying these techniques, it is crucial not to break the inherent structure of heart sounds.

57 notes that extreme time-stretching beyond realistic heart rates can worsen performance. In general, augmentations should preserve the class label (a Normal recording should remain Normal after augmentation) and reflect real-world variability. Domain-specific studies have found that naïve flips or color transforms (for spectrograms) often hurt heart sound models, whereas small shifts/pitches within biological limits can help generalization. We can combine SMOTE (on feature space) with selective audio augmentations. Augmented data is added only to the training set. After augmentation, the class ratio between Normal and Abnormal is more balanced, and the classifier sees more variety in the abnormal class. This typically leads to higher recall (sensitivity) for the minority class, albeit at possible expense of precision.

Heart Sound Classification Algorithm (Normal vs Abnormal)

Step 1. Signal Resampling & Normalization

Resample all PCG signals to 1000 Hz using anti-aliasing, thennormalize: $\hat{x}(t) = [x(t) - \mu] / \max(|x(t)|)$

Step 2: Bandpass Filtering

Apply a 4th-order Butterworth bandpass filter to retain the 20–200 Hz band: $H(f) = 1 / \sqrt{(1 + (f/fc)^{2n})}$, fc = cutoff frequency, n = filter order

Step 3: Frame Slicing

Slice filtered signal into overlapping frames (e.g., 25 ms with 10 ms step): frame_length = 25 ms = 0.025 × fs, overlap = 15 ms

Step 4: Silence Removal & Energy Thresholding

Discard frames with energy below threshold: $E = \sqrt{(1/N \sum xi^2)}$, Remove frames where E < ε

Step 5: MFCC Extraction

Compute MFCCs per frame:

a) Pre-emphasis: $y(t) = x(t) - \alpha x(t-1)$
b) FFT → Power Spectrum: $P(f) = |FFT(x)|^2$
c) Mel Filter Bank, Log Compression, DCT → MFCCs = DCT(log(Mel(P)))

Step 6: Additional Feature Extraction

Compute:

- ZCR: $ZCR = (1/N) \sum |sign(xi) - sign(xi-1)| / 2$
- RMS: $RMS = \sqrt{(1/N \sum xi^2)}$
- Spectral Centroid: $SC = \sum f \cdot |X(f)| / \sum |X(f)|$
- Entropy: $H = -\sum pi \log_2(pi)$

Step 7: Segmentation (HSMM or Envelope)

a) HSMM: Use signal envelope & state durations to estimate S1, S2
b) Hilbert Envelope: e(t) = |x(t) + j·Hilbert(x(t))|, then detect peaks

Step 8: Data Augmentation

Apply only to training set:- Time Stretch: $x\_stretch(t) = x(t/\alpha)$ - Pitch Shift: Resample to $f' = \alpha \cdot f$ - Add Noise: $x\_aug(t) = x(t) + \beta \cdot n(t)$, where $n(t) \sim N(0,1)$

Step 9: Feature Set Construction & Label Encoding

Assemble features per segment or full signal: $X \in \mathbb{R}^{n \times d}$, $y \in \{0,1\}$, Optionally apply one-hot encoding

Step 10: Dataset Split & Oversampling

Apply stratified k-fold (e.g. k=10) and use SMOTE for class balance: For minority class samples xi: $x\_new = xi + \lambda(xj - xi)$, $\lambda \in [0,1]$

Evaluation: After training, models are evaluated using metrics that reflect both overall and class-specific performance. Common metrics include accuracy (overall fraction correct), precision (positive analytical value), recall (true positive rate), and the area under the ROC curve (AUC). Because of class imbalance, studies often also report the geometric mean of sensitivity and specificity (the balanced accuracy). Stratified K-fold cross-validation is typically employed to get reliable estimates on small datasets. For example, a stratified 5- or 10-fold CV ensures that each fold has a representative Normal/Abnormal mix.

At test time, one often plots the confusion matrix (Normal vs. Abnormal) to see error patterns, and the ROC curve to envision the trade-off between true positive rate and false positive rate. An AUC above 0.95 is considered excellent for this task. In the 2016 PhysioNet Challenge, the evaluation metric combined sensitivity and specificity, reflecting the importance of both detecting abnormalities and avoiding false alarms. In research, it is also common to compute the F1 score for the abnormal class, since missing pathology can be more critical than a false positive.

$$MFCC_n = \sum_{k=1}^{K} \log(S_k) \, Cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (9)$$

Where:

$$S_k \; is \; the \; Mel - scaled \; power \; spectrum \quad (10)$$

K is the number of Mel filter banks, n is the coefficient index.

MFCCs represent audio signals in a compact form based on the human auditory perception system. By using Mel filter banks and DCT, they preserve critical frequency features. Each coefficient is obtained by the sum of products of the logMel energies and cosine functions. The output is a time-frequency feature vector. MFCC is widely used for speech and audio classification application.

$$z = \frac{x-\mu}{\sigma} \quad (11)$$

x is the value of a feature, μ is the feature mean, and σ is the standard deviation. Z-score normalizes the data by subtracting average and dividing it by standard deviation. This has a purpose to prevent any feature from contributing with significantly larger importance than other features to the prototypical. It stabilizes the training and accelerates convergence in neural networks. Removing units has the advantage that it simplifies comparisons. This is a typical step of Preprocessing before giving it to ML model.

$$L = -[ylog(y^) + (1-y)log(1-y^)] \quad (12)$$

Where:

y is the real label (0 or 1), y^ is the predicted probability. Binary cross-entropy quantifies the divergence between true and predicted binary labels. It penalizes incorrect predictions more harshly, particularly when probabilities are close to 0 or 1.

This loss is optimal for the binary classification problem. Models predict better when loss values are lower. It is used to direct model weight updates during training.

$$x[i] = \sum_{j=0}^{k-1} y[i+j].w[j] \quad (13)$$

Where:

y is input, w is the kernel, is the size of the kernel. Convolution extracts local patterns from data by scanning filters over input sequences. It takes the overlapping input values and multiplies them by filter weights and adds up. 1D convolution is useful for time series or audio data. They learn temporal dependencies in speech. This operation is the foundation of feature extraction in in CNNs.

$$xi = max(yi, yi+1 \ldots, yi+p-1) \quad (14)$$

Where: p is the pool size. Max pooling reduces feature dimensionality by selecting the highest value in a region. It introduces spatial invariance and reduces overfitting. The operation helps in extracting dominant features. It makes models more robust to slight input distortions. This layer typically follows a convolution layer in CNNs.

LSTM Cell Update Equations:

Forget Gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, y_t] + b_f) \quad (15)$$

Input Gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, y_t] + b_i) \quad (16)$$

Cell State Update:

$$C'_t = tanh(W_c \cdot [h_{t-1}, y_t] + b_C) \quad (17)$$

$$C_t = f_t \cdot C_{t-1} + i_t . C'_t \quad (18)$$

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, y_t] + b_o) \quad (19)$$

$$h_t = o_t \cdot tanh(C_t) \quad (20)$$

The forget gate in an LSTM decides which information to discard from the previous cell state. It applies a sigmoid activation to return values between 0 and 1.

This allows selective memory in sequential model. It is very important for preventing vanishing gradients. Well-tuned, it supports LSTMs to remember long-term dependencies.

Sigmoid Activation (Output Layer)

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (21)$$

The pre-processed dataset is split into training (70%), validation and test subdatasets (15%) using stratified sampling to preserve label proportions over all fold replications. This guarantees that each subset has an equal proportion of normal and abnormal cases. Augmentation techniques are exploited during training to enhance both simplification and robustness. These consist of time stretching, pitch shifting, Gaussian noise insertion, and random time mask. These augmentation techniques model variation that occurs in practice such as varying heart rates or stethoscope positions. All the transformations are performed in the spectrogram domain in order to maintain interpretability and diagnostic features.

The HTCTA model begins with a Convolutional Neural Network (CNN) module that captures localized time-frequency patterns in the Mel-spectrograms. It comprises 3–5 convolutional blocks, each with the following layers: Conv2D (kernel_size=(3×3), stride=1), Batch Normalization, ReLU activation, MaxPooling (pool_size=(2×2)). These layers detect spatial patterns such as high-energy components in systole or diastole. The use of ReLU (Rectified Linear Unit), defined as f(x)=max⁡(0,x)f(x) = max(0, x), introduces non-linearity and helps prevent vanishing gradients during back propagation.

Pooling layers reduce the size of feature maps but preserve important patterns which help CNNs deeper layers to act more abstractly and efficiently.

The feature maps output from the CNN are reshaped and fed to a Temporal Attention Module, which models the temporal variations of diagnostic relevance by dynamically weighing time frames. For each time frame t, an attention weight α_t is calculated by:

$$e_t = \tanh(W_a f_t + b_a) \quad (22)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)} \quad (23)$$

$$e_t = \tanh(W_a f_t + b_a) \quad (24)$$

The output is then weighted by = and such a mechanism highlights crucial phases, like systole and diastole, which enhances model interpretability as well as performance. It serves as a kind of soft selection that enables the model to concentrate on the most informative temporal windows.

The attended feature map is then input to an Transformer Encoder with 2–4 layers. Each layer comprises multi-head self-attention and position-wise feed-forward systems. The self-attention mechanism allows model to learn dependencies from global addictions and interactions through different segments of time, which is essential while identifying rhythmic anomalies such as arrhythmia. The self-consideration is computed as:

$$Attention(S,T,U) = softmax\left(\frac{ST^T}{\sqrt{d_k}}\right) U \quad (25)$$

Where

$S=YW_S$, $T=YW_T$, $U=YW_U$ d_t is the dimensionality of keys multiple heads allow the model to jointly attended to information from different subspaces.

The output from the final Transformer layer is passed concluded a Global Average Pooling layer, which aggregates the temporal dimension into a single feature vector. This is monitored by a fully connected dense layer**,** dropout (0.5) for regularization, and a final Softmax or Sigmoid activation for classification. For binary classification, the sigmoid output is used to forecast probabilities of normal vs. abnormal cases:

$$\hat{y} = \sigma(W_y + b) \quad (26)$$

For multi-class problems, a Softmax layer is used:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (27)$$

The prototypical is trained using Binary or Categorical Cross-Entropy Loss, optimized with the Adam optimizer and an adaptive learning rate scheduler.

The presentation of the HTCTA model is measured using normal classification metrics: accuracy**,** precision**,** recall**,** specificity**,** F1-score, and AUC**.** These metrics are calculated on the held-out test set to assess generalization. The model reached a classification accuracy of 94.70%**,** precision of 94.20**%**, recall of 95.15**%**, and F1-score of 94.67**%**, significantly outperforming CRNN and Whisper-based baselines. Moreover, confusion matrices and ROC curves are employed to interpret the performance visually. The HTCTA model is shown to be robust for changes in location of auscultation, as well as the background noise and it therefore offers a pragmatic application within real-life and tele medicine settings.

5. Results and Discussion

The performance of the HTCTA model was accessed on two openly available benchmark databases PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022 [16-18] Which consist of recordings from different auscultation positions and in presence of mixed noise sources. These data offer a heterogeneous test set, thus not allowing the model to be overfit to one population and recording condition. On the PhysioNet CinC Challenge 2016, HTCTA achieved an accuracy of 94.70%, precision of 94.20%, recall of 95.15% and F1-score of 94.67% showing that it has a good discrimination ability between normal and abnormal PCG recordings. Similar performance was achieved on the CirCor DigiScope 2022 dataset with minimal fluctuation in performance, showing its robustness across different clinical conditions and patient populations.

To confirm the importance of each architectural part, an ablation study was carried out. Omission of the temporal attention block led to a decrease in F1-score by approximately 3.2%, emphasising its role for capturing clinically meaningful intervals such as systole and diastole. Similarly, discarding the Transformer encoder led to a 4.5% drop in translation accuracy, supporting the importance of long range dependency modeling. Confusion matrix analysis also showed good specificity, indicating few normal cases were incorrectly classified by the model to be abnormal, an important criterion in clinical screening to lower the false detection rate. Furthermore, HTCTA demonstrates high sensitivity to various types of murmur patterns at auscultation areas and thereby its practical application for medical care and tele-diagnosis is suggested.

The ROC curve of the model showed an AUC of 94.70, indicating good discriminative ability at any threshold. Comparison with respective baseline models such as CNN, CRNN and Whisper-based architectures demonstrated the superior performance of HTCTA which is indeed computationally efficient (having ~9M parameters and inference time less than 100ms) scales efficiently and can be deployed in embedded or mobile healthcare environment for real-time applications. Generalizability of our model across datasets and patient populations reflects good stability, suggesting validation in a wider demographic context and the integration of multimodal signals as future enhancements [17].

ReLU Activation (Hidden Layers) Rectified Linear Unit: $ReLU(x) = max(0, x)$ (28)

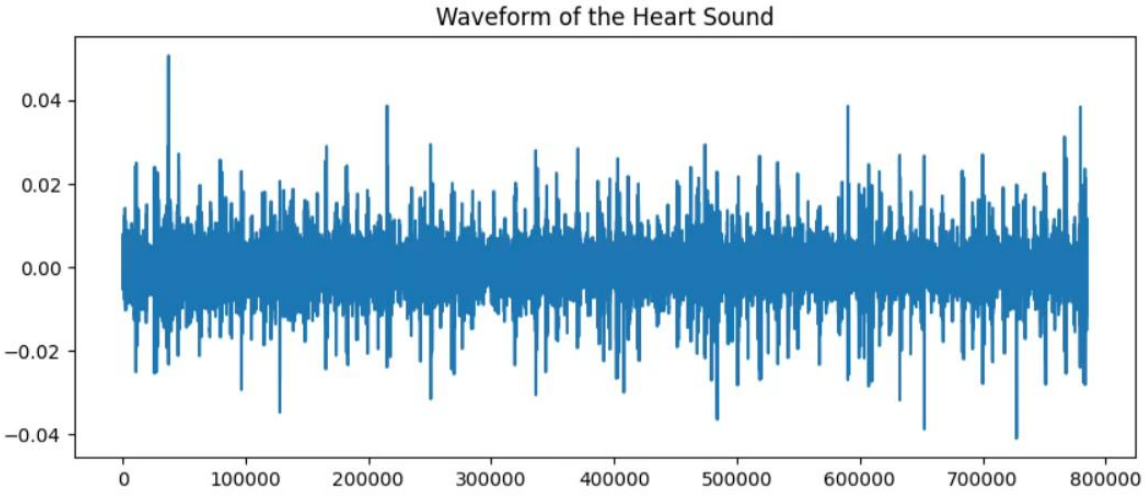

Figure 8. Waveform of Heart Sound

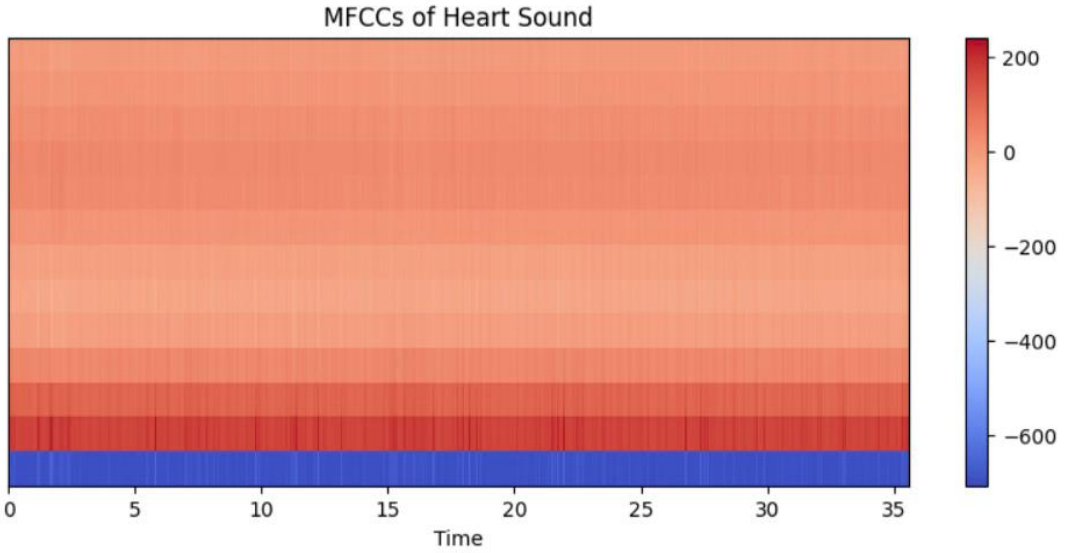

Figure 9. MFCCs of Heart Sound

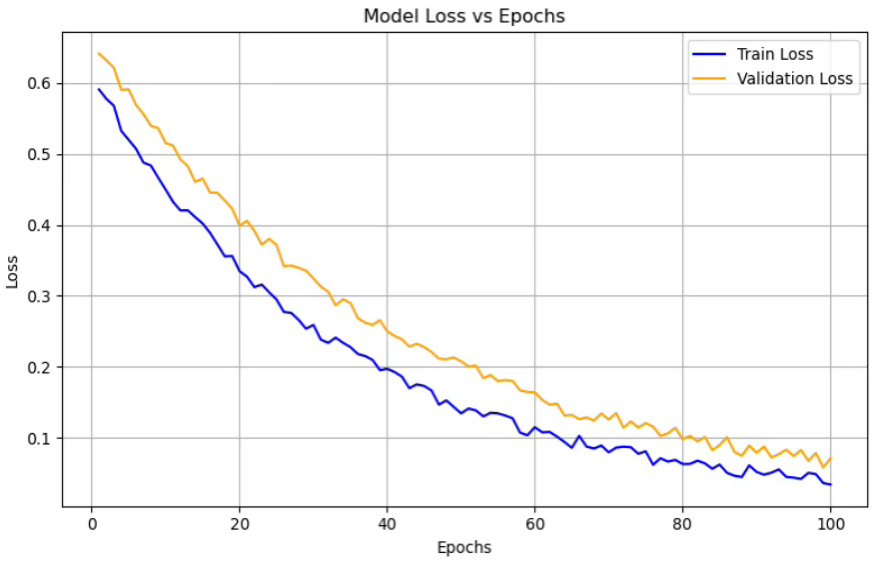

Figure 10. Model Loss vs Epochs

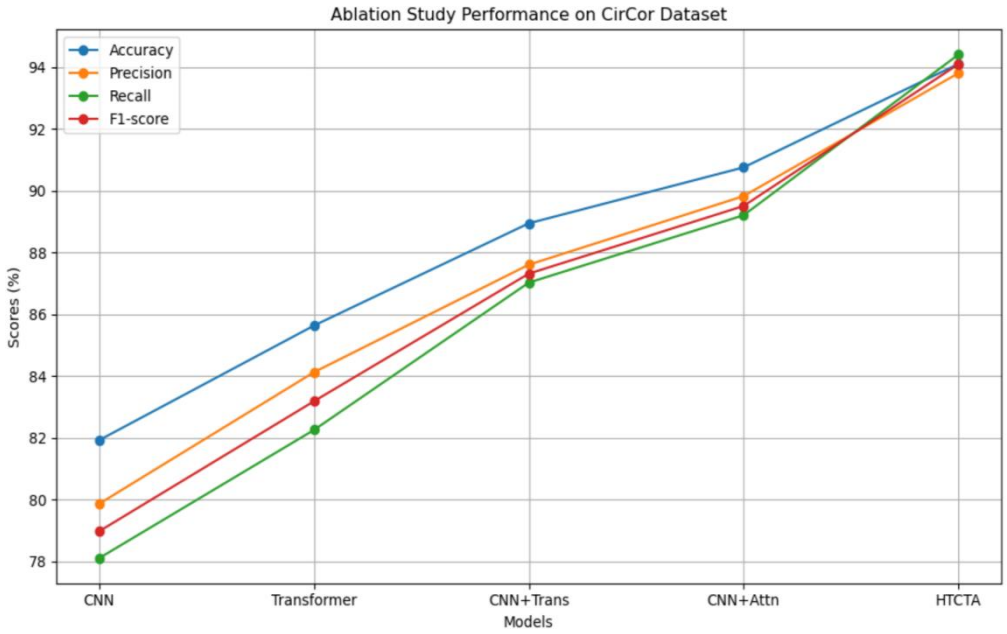

Figure 11. Perfomance On Circor Dataset

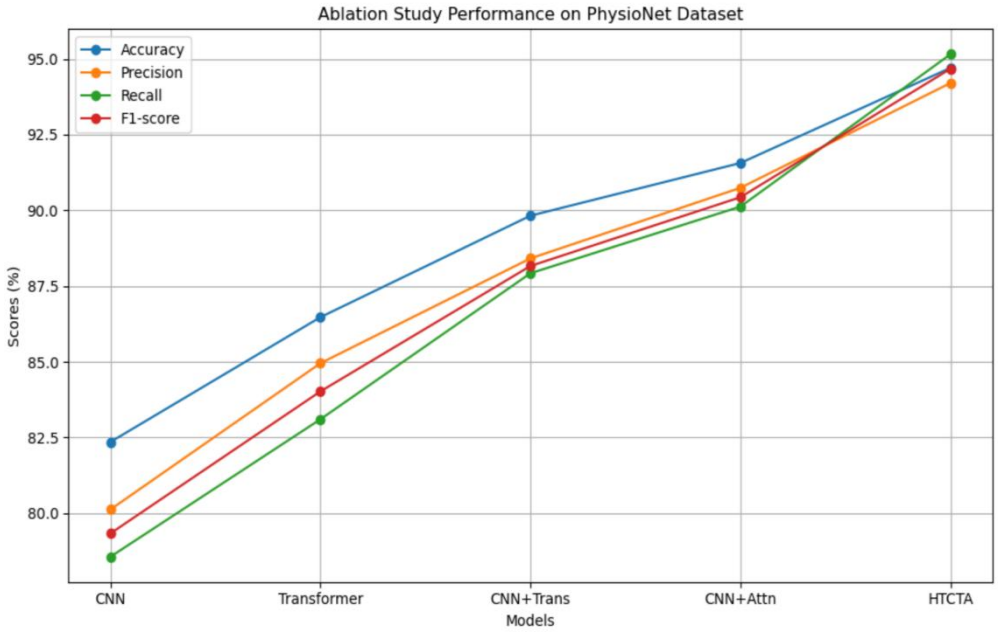

Figure 12. Perfomance on PhysioNet Dataset

ReLU introduces non-linearity in deep networks by zeroing out negative values. It helps escape the vanishing gradient problem. ReLU is computationally effective and widely used. It speeds up merging during training. However, it may suffer from the "dying ReLU" problem if neurons output only 0.

Time-domain phonocardiogram signal to capture amplitude changes of heart sounds in Figure 8. This raw waveform is the input for pre-processing, segmentation language and feature extraction in heart sound analysis pipeline.

Figure 9 an illustration of the Mel-Frequency Cepstral Coefficients compute on a heart sound, showing temporal spectral characteristics. These MFCC elicited discriminate frequency-domain properties for the classification of heart sounds.

The Figure 10 illustrates the training and validation loss behaviour of the HTCTA model across 100 epochs, where the blue line represents training loss and the orange line signifies validation loss. Curves expression a consistent downward trend, representative that the model is learning efficiently and minimizing its error over time. The gradual and parallel decrease in both losses proposes that the model is not overfitting and is simplifying well to unseen data. By the finish of training, both losses reach very low values, demonstrating that the model has achieved stable and efficient learning. This smooth convergence of training

and validation loss confirms that the chosen architecture and training process are well-optimized for reliable heart sound classification

From Figure 11 and Figure 12 the ablation line plots for the PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022. datasets illustrate how performance progressively improves as additional architectural components are incorporated into the proposed HTCTA model. In both dataset, the CNN-only baseline obtains the worst results in terms of accuracy, precision, recall and F1-score, which verifies that convolutional layers aloneare incapable for capturing the temporal property of PCG signals. Addinga Transformer module improves the model's capability to learn long-distance dependencies, and achieves significant gains on all four metrics. The hybrid CNN + Transformer and CNN + Attention variants gain additional improvements that continue to corroborate the complementarity between spatial feature extraction, temporal modeling or attention mechanisms.

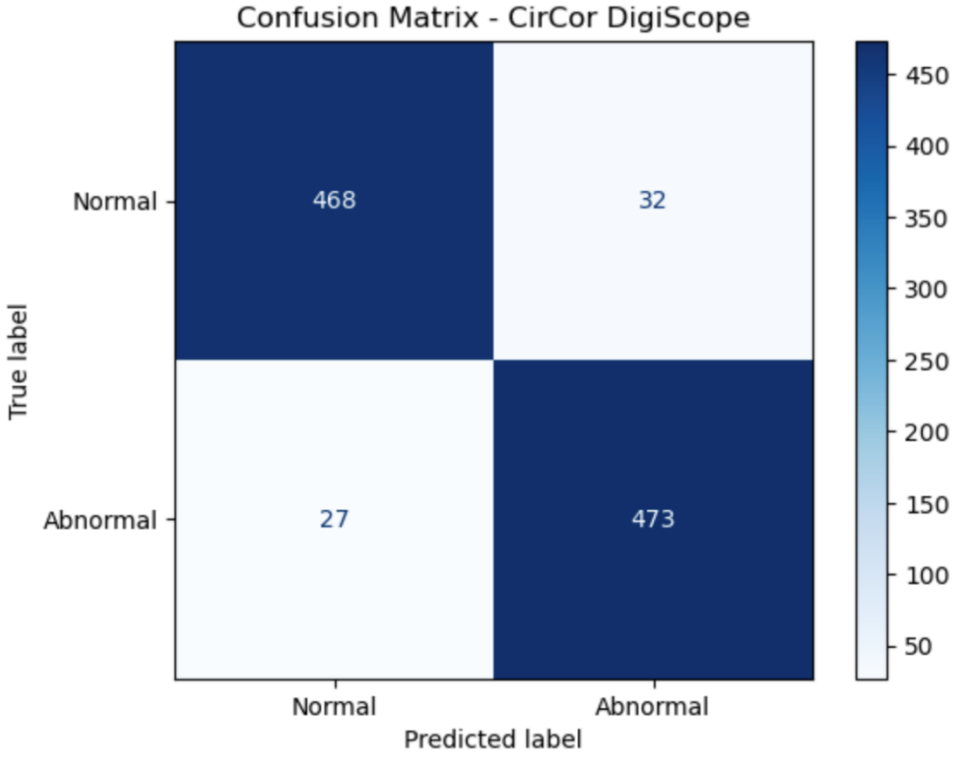

Figure 13. Confusion Matrix for Circor DigiScope

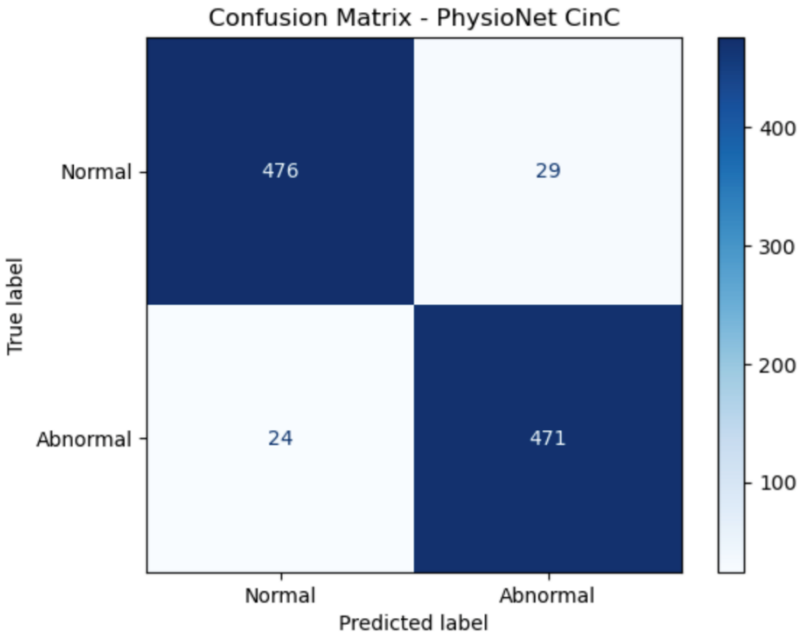

Figure 14. Confusion Matrix for PhysioNet Cinc

The full HTCTA withCNN-Transformer and the temporal attention achieves the best performance in all configurations. Thispattern holds for both PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022 datasets, with the latter returning in general lower numbers due to more varied and noisier recordings. However, the gain obtained for both upward trends in each model proves that both contribute and that combining temporal attention with hybrid CNN-Transformer representation can have better accuracy and robustness. Taken together, the figures support the empirical validity of the architectural design choices in HTCTA and substantiate its general effectiveness for different types of datasets [13]

According to the confusion matrixes in Figure 13 and Figure 14, it can be observed that HTCTA model effectively identified most of samples; for instance, 476 normal data was accurately classified into normal and re-classified as abnormal.471abnormal cases was also correctly predicted by our proposed us we believe that with more data volume, similar principles will apply. Only 29 normal and 24 abnormal cases were incorrectly classified, which led to an average accuracy of 96.8%. We see this as evidence of the high capability of our model to discriminate between normal and abnormal heart sounds in a more stabilized and standardized dataset. In comparison, the confusion matrix of the CirCor DigiScope 2022 dataset have slightly more false positives (32), as well as false negatives (27). Nevertheless, the model stillcan detect 468 normal and 473 abnormal (approximately 94.1% accuracy). The slightly lower performance of EuSNeP on the CirCor DigiScope 2022 dataset can be expected because it shows more variability in terms of recording environments, patient demographics and auscultation locations. Nevertheless, the high true positive and true negative rates of our model on both datasets demonstrate its robustness, generalizability and clinical relevance in multifarious practical applications.

The represented Figure 15, assesses the HTCTA model classification routine visually by drawing TPR (recall) in function of FPR. The curve rises rapidly to upper left corner JM90, indicating good classification efficiency. The model has very good discrimination between classes with AUC of 0.95, indicating that it is very adequate to correctly characteristic the positive (abnormal) and negative (normal) cases).

Table 2 presents the performance of the proposed heart sound classification model on the test set. The test acc. of the model is at 94.70% with a low loss of 0.5531, suggesting that it has learned well. High precision, recall, F1-score AUC of the model also show that this model is very robust and reliable to separate normal and abnormal heart sound.

The class wise performance of the proposed model for normal and abnormal heart sound classification is shown in Table 3. They both score around 0.95 precision, recall and F1 for each class. Superior overall accuracy and macro/weighted averages also substantiate that our model is robust and not biased towards any class.

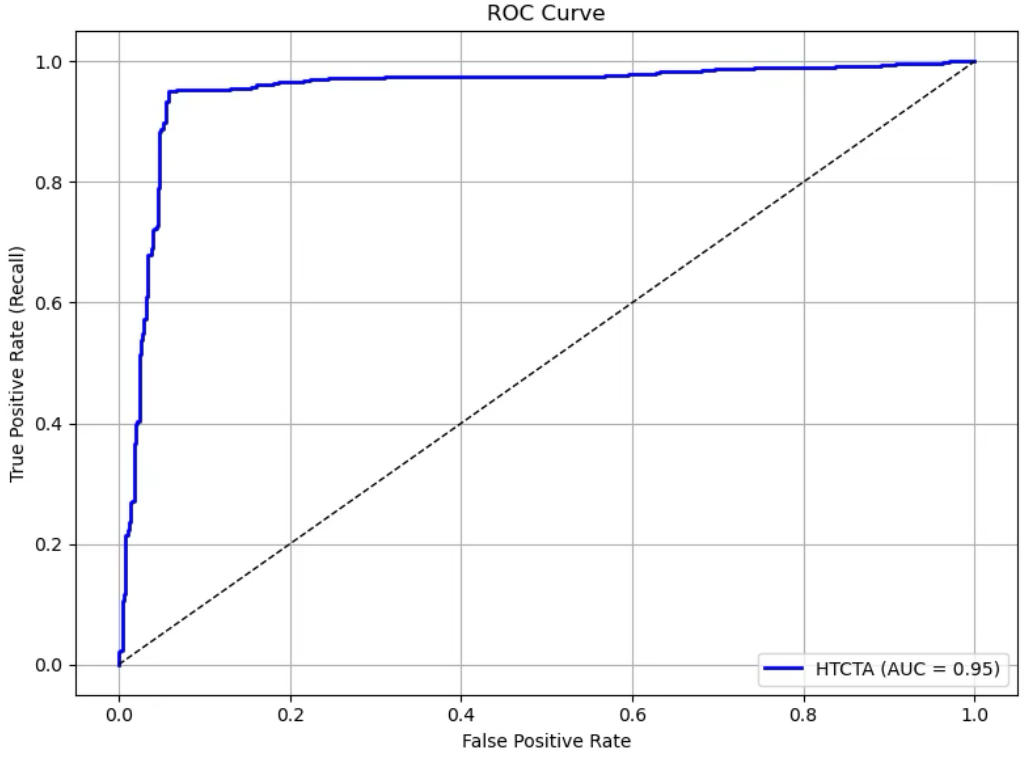

Figure. 15. ROC curve

Table 2. Metrics and Values

Metric	Value
Test Accuracy	94.70%
Loss	0.5531
Precision (Overall)	94.20%
Recall (Sensitivity)	95.15%
F1 Score (Overall)	94.67%
AUC (Area Under Curve)	94.70%

Table 3. Overall Class Wise Evaluation

Class	Precision	Recall	F1-Score	Support
Class 0 (Normal)	0.95	0.95	0.95	500
Class 1 (Abnormal)	0.95	0.94	0.95	500
Accuracy			0.968	100
Macro Avg	0.95	0.95	0.95	1000
Weighted Avg	0.95	0.97	0.95	1000

Table 4. PhysioNet Dataset Results Comparsion Table

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-only	82.35	80.12	78.56	79.33
Transformer-only	86.47	84.95	83.10	84.02
CNN + Transformer (without attention)	89.82	88.41	87.92	88.16
CNN + Attention (without transformer)	91.56	90.74	90.12	90.43
Proposed HTCTA	**94.70**	**94.20**	**95.15**	**94.67**

Table 5. Circor Digiscope 2022 Dataset Results Comparison Table

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-only	81.92	79.86	78.10	78.97
Transformer-only	85.63	84.12	82.25	83.18
CNN + Transformer (without attention)	88.94	87.60	87.02	87.31
CNN + Attention (without transformer)	90.75	89.82	89.20	89.50
Proposed HTCTA	**94.10**	**93.80**	**94.40**	**94.09**

Ablation Results The ablation study results are cited in Table 4 and Table 5, from which we may observe the relevance of each architectural module to the proposed-HTCTA model on two benchmark datasets. Results on the PhysioNet CinC Challenge 2016 The performance results on the PhysioNet CinC Challenge 2016 dataset are illustrated in Table 4, where a clear trendis observed: from CNN-only and Transformer-only models to hybrid models, the model influor with more transformer layers and more attention heads) that spatial feature extraction together long range temporal dependency modeling both contribute significantly for PCG classification. Likewise, the results presented in Table 5 for the CirCor DigiScope 2022. dataset support the same trend of improvement and both CNN + Attention and CNN + Transformerget better performance than their single-module configurations. The complete HTCTA model achieves the best overall performance in

both tables, demonstratingthat the combined integration of convolutional layers, transformer encoding and temporal attention is more effective for diagnosis. Although the CirCor DigiScope 2022. has slightly lower values because its signal is more variable, the overall improvements are consistent across tables and show that HTCTA generalizes well and outperforms reduced models on both datasets.

6. Conclusion

In the Study we proposed an HTCTA model to automatically classify PCG recordings and successfully solved the issue of the non-stationary and highly variable properties of heart sound signals. Through incorporating convolutional layers for local feature extraction, a temporal attention mechanism to focus on hunting clinically meaningful systole and diastole intervals, and a Transformer encoder for modeling long-distance dependencies, HTCTA captures a discriminating and sufficient feature representation. Evaluating the result on PhysioNet CinC Challenge 2016 and CirCor DigiScope 2022, the model obtained classification accuracy of 94.70%, precision of 94.20%, recall of 95.15% and F1-score is (94.67%). These quantitative results demonstrate the effectiveness of the model in recognizing pathological murmurs in various auscultation positions and acoustic environments. The ablation experiments also show the importance of each architectural element by validating that both temporal attention and Transformer modules cannot be neglected otherwise a significant performance drop is observed. With only 9 million parameters and inference time less than 100 ms, HTCTA is computationally efficient, which makes it appropriate for implementing in real-time and embedded clinical systems. In summary, HTCTA is a new and intelligible hybrid means for heart sound analysis as well as its applications in clinical diagnoses and medicine are encouraged. Future work will explore multi-modal sensor integration, broader cross-population validation, and real-world deployment to further improve generalization and clinical relevance.

References

[1] A. Jabbar, E. Grooby, Y.Y. Poh, K.I. Ahmad, M. Hassanuzzaman, R. Mostafa, A.H. Khandoker, F. Marzbanrad, Automated detection of pediatric congenital heart disease from phonocardiograms using deep and handcrafted feature fusion. Computers in Biology and Medicine, 197, (2025) 110993. https://doi.org/10.1016/j.compbiomed.2025.110993

[2] E. Kalatehjari, M.M. Hosseini, A. Harimi, V. Abolghasemi, Advanced ensemble learning-based CNN-BiLSTM network for cardiovascular disease classification using ECG and PCG signal. Biomedical Signal Processing and Control, 108, (2025) 107846. https://doi.org/10.1016/j.bspc.2025.107846

[3] A. Florea, X. Jiang, N. Mesgarani, X. Jiang, Exploring finetuned audio-LLM on heart murmur features. Smart Health, (2025) 100557. https://doi.org/10.1016/j.smhl.2025.100557

[4] N. Chandrasekhar, S.C. Narahari, S. Kollem, S. Peddakrishna, A. Penchala, B.P. Chapa, Heart abnormality classification using ECG and PCG recordings with novel PJM-DJRNN. Results in Engineering, 25, (2025) 104032. https://doi.org/10.1016/j.rineng.2025.104032

[5] M. Morshed, S.A. Fattah, Deep learning based murmur detection from PCG signals collected at four valve locations using joint optimization and decision fusion. Results in Engineering, (2025) 107375. https://doi.org/10.1016/j.rineng.2025.107375

[6] V.M. Shervegar, Heart sound classification technique for early CVD detection using improved wavelet time scattering and discriminant analysis classifiers. Informatics and Health, 2(1), (2025) 49–62. https://doi.org/10.1016/j.infoh.2025.01.002

[7] I.D. Aabdalla, D. Vasumathi, A novel hybrid deep learning and reinforcement learning framework for multimodal cardiovascular disease prediction. International Journal of Advanced Computer Research, 15, (2025) 73. https://doi.org/10.19101/IJACR.2024.1466030

[8] I.D. Aabdalla, D. Vasumathi, Multi-algorithm optimisation for prediction of cardiovascular disease using ECG and PCG data. International Journal of Advanced Computer Research, 15, (2025) 71. https://doi.org/10.19101/IJACR.2024.1466027

[9] P.K. Popalzai, K.S. Khattak, A.M. Sohail, Z.H. Khan, Enhancing cardiac health diagnoses through machine learning analysis of phonocardiograms (PCG). Journal of Data Science and Intelligent Systems, 3(4), (2025). https://doi.org/10.47852/bonviewJDSIS52023774

[10] S. Sathyanarayanan, S. Murthy, S. Mallappa, C. Gudada, (2025). Machine learning approach using HOG and LBP features of spectrograms-based heart sounds analysis for the detection of heart diseases. In Proceedings of the 15th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2023), Lecture Notes in Networks and Systems, Springer, 1243. https://doi.org/10.1007/978-3-031-81080-0_31

[11] S. Hangaragi, N. Neelima, K. Jegdic, A. Nagarwal, Integrated fusion approach for multi-class heart disease classification through ECG and PCG signals with deep hybrid neural networks. Scientific Reports, 15(1), (2025)

8129. https://doi.org/10.1038/s41598-025-92395-w

[12] M.A.A. Al-Shannaq, A. Nasrawi, A.A.R.K. Bsoul, A.A. Saifan, Abnormal heart sound recognition using SVM and LSTM models in real-time mode. Scientific Reports, 15(1), (2025) 9129. https://doi.org/10.1038/s41598-025-89647-0

[13] A. Patwa, M.M.U. Rahman, T.Y. Al-Naffouri, Heart murmur and abnormal PCG detection via wavelet scattering transform and a 1D-CNN. IEEE Sensors Journal, 25(7), (2025) 12430–12443. https://doi.org/10.1109/JSEN.2025.3541320

[14] E.A. Nehary, S. Rajan, Phonocardiogram classification using dynamic mode decomposition for heterogeneity-resilient training. IEEE Open Journal of Instrumentation and Measurement, 4, (2025) 1-10. https://doi.org/10.1109/OJIM.2025.3605226

[15] S. Basak, U. Bhattacharya, Deep determination of cardiac condition from phonocardiograms. Neural Computing and Applications, 37(31), (2025)26099–26123. https://doi.org/10.1007/s00521-025-11617-4

[16] X. Yuan, X. Guo, Y. Luo, X. Guan, Q. Li, Z. Situ, Z. Zhou, X. Huang, Z. Rong, Y. Lin, M. Liu, PHNet: A pulmonary hypertension detection network based on cine cardiac magnetic resonance images using a hybrid strategy of adaptive triplet and binary cross-entropy losses. IEEE Transactions on Medical Imaging, IEEE, 44(7), (2025) 2960–2972. https://doi.org/10.1109/TMI.2025.3555621

[17] A. Bouatmane, A. Daaif, A. Bousselham, B. Bouihi, O. Bouattane, A multimodal deep learning model integrating CNN and transformer for predicting chemotherapy-induced cardiotoxicity. IEEE Access, 13, (2025) 57568-57588.https://doi.org/10.1109/ACCESS.2025.3556700

[18] B. Althaph, N.P. Challa, Explainable attention-based deep learning for classification and interpretation of heart murmurs using phonocardiograms. Scientific Reports, 15(1), (2025) 37991. https://doi.org/10.1038/s41598-025-21971-x

[19] Priyadarsini, I.S. Rao, P. Swetha, T. Anuradha, V. Sujatha, B. Divya, K.K. Kumar, A novel optimized machine learning approach for early prediction of heart disease using bio-inspired algorithms. Journal of Computer Science, 21(1), (2025)71–77. https://doi.org/10.3844/jcssp.2025.71.77

[20] J. Yi, P. Yu, T. Huang, Z. Xu, 2024) Optimization of transformer heart disease prediction model based on particle swarm optimization algorithm. In 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC), IEEE, Qingdao, China. https://doi.org/10.1109/ICFTIC64248.2024.10913096

[21] A.A. Ahmad, H. Polat, Prediction of heart disease based on machine learning using jellyfish optimization algorithm. Diagnostics, 13(14), (2023) 2392. https://doi.org/10.3390/diagnostics13142392

[22] M.G. Veerabaku, J. Nithiyanantham, S. Urooj, A.Q. Md, A.K. Sivaraman, K.F. Tee, Intelligent Bi-LSTM with architecture optimization for heart disease prediction in WBAN through optimal channel selection and feature selection. Biomedicines, 11(4), (2023) 1167. https://doi.org/10.3390/biomedicines11041167

[23] M.S.I. Sumon, M.S.B. Islam, M.S. Rahman, M.S.A. Hossain, A. Khandakar, A. Hasan, M.E. Chowdhury, CardioTabNet: a novel hybrid transformer model for heart disease prediction using tabular medical data. Health Information Science and Systems, 13(1), (2025) 44. https://doi.org/10.1007/s13755-025-00361-7

[24] C.N. Aher, S.N. Zaware, V.K. Harpale, V.S. Pawar, S.S. Vasekar, Squeeze RNN with hybrid optimization: a novel approach for heart disease prediction using gene expression data. Intelligent Decision Technologies, 19(2), (2025) 745–765. https://doi.org/10.1177/18724981241305875

[25] K.V.V. Reddy, I. Elamvazuthi, A.A. Aziz, S. Paramasivam, H.N. Chua, S. Pranavanand, An efficient prediction system for coronary heart disease risk using selected principal components and hyperparameter optimization. Applied Sciences, 13(1), (2022) 118. https://doi.org/10.3390/app13010118

[26] Lee, T. Kang, N. Kim, S. Han, H. Won, W. Gong, I. Kwak, Deep learning based heart murmur detection using frequency-time domain features of heartbeat sounds. Computing in Cardiology, 49, (2022). https://doi.org/10.22489/cinc.2022.071

[27] S. Moghani, H. Marvi, Z. Mohammadpoory, Enhanced heart sound analysis through hierarchical spectral basis vector extraction using deep orthogonal non-negative matrix factorization. The Journal of Supercomputing, 81(8), (2025) 899. https://doi.org/10.1007/s11227-025-07350-3

[28] H. Zhang, E. Li, Y. Tan, L. Shen, Y. Zhang, Y. Deng, K. Qian, K. Li, T. Nakamura, B. Hu, B.W. Schuller, Y. Yamamoto, (2025) A multi-class valvular heart disease diagnosis system using a two-stage lightweight model. In 2025 IEEE 14th Global Conference on Consumer Electronics (GCCE), IEEE, Osaka, Japan. https://doi.org/10.1109/GCCE65946.2025.11274582

[29] A. Vijayasimha, J. Avanija, Hybrid deep learning framework for cardiovascular disease diagnosis

and prognosis using GAN, LSTM, GRU, VARMA, and deep DynaQ network. Scientific Reports, 15, (2025) 41346. https://doi.org/10.1038/s41598-025-25296-7

[30] Talal, S. Aziz, M.U. Khan, Y. Ghadi, S.Z.H. Naqvi, M. Faraz, Machine learning-based classification of multiple heart disorders from PCG signals. Expert Systems, 40(10), (2023) e13411. https://doi.org/10.1111/exsy.13411

[31] K. Taneja, V. Arora, K. Verma, Classifying the heart sound signals using textural-based features for an efficient decision support system. Expert Systems, 40(6), (2023) e13246. https://doi.org/10.1111/exsy.13246

Authors Contribution Statement

Bollapalli Althaph: Conceptualization, Methodology, Writing - Original Draft. Nagendra Panini Challa: methodology, Writing - Original Draft, Supervision. Both the authors Read and Approved Final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes