Explainable and Accurate Valvular Heart Disease Screening Using Spectrogram-Based 2D-CNNs on PCG Signals

Talha Ahmed Siddiqui[1], Sania Akhtar[1,2(✉)], Muhammad Taha Mansoor[1], Badee Baltow[3], Sadaq-Ur Rehman[4(✉)], and Noor Ul Huda[1]

[1] Faculty of Computer Science and Artificial Intelligence, Department of Creative Technology, Air University, Islamabad, Pakistan

[2] Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), Topi 23460, Pakistan
sania.akhtar@au.edu.pk

[3] King Abdullah Medical City, Makkah, Saudi Arabia

[4] School of Sciences, Engineering and Environment, University of Salford, Manchester M5, UK
s.rehman15@salford.ac.uk

Abstract. Despite the recent advancements in artificial intelligence the cardiovascular diseases remain one of the leading causes of mortality worldwide, emphasizing the need for reliable and accessible diagnostic techniques. For the detection of the cardiac abnormalities, phonocardiography (PCG) which records heart sounds during the cardiac cycle, provides a non-invasive and cost-effective method. Conversely, manual interpretation of PCG signals depends heavily on clinician expertise and is often difficult by co-occurring pathologies, signal noise, and class imbalance in clinical datasets. This research focuses on an end-to-end deep learning pipeline for automated multi-label analysis of valvular heart diseases using PCG recordings. BMD-HS dataset containing 872 recordings sampled at 4000 Hz used for experimentations. The developed system implements a dual-stream learning strategy: handcrafted feature extraction combined with XGBoost for interpretability, and deep learning representations using one-dimensional and two-dimensional convolutional neural networks for improved predictive performance. Experimental results establish that altering raw PCG signals into Mel-Spectrogram representations and then it processes them with a 2D-CNN architecture achieving the best performance, with an AUC of 0.92 and an F1-score of 0.94 for the normal class. Furthermore, Explainable Artificial Intelligence techniques are implemented to confirm that the model focuses on medically significant spectral features. The developed framework provides a robust and low-cost solution for automated cardiac screening in resource-restricted healthcare environments.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16766, pp. 135–152, 2027.
https://doi.org/10.1007/978-3-032-30539-8_9

Keywords: Phonocardiography · Valvular heart disease detection · CNN · Mel-spectrogram analysis · Multi-label classification · Explainable artificial intelligence

1 Introduction

Cardiovascular diseases (CVDs) remain one of the leading causes of death globally, necessitating reliable and affordable diagnostic methods [1,2]. Among these, phonocardiography (PCG) captures acoustic signals produced by cardiac valve activity and associated hemodynamics [3,4]. In a healthy cardiac cycle, two primary sounds are observed: S1, generated by mitral and tricuspid valve closure at systole onset, and S2, generated by aortic and pulmonary valve closure at diastole onset. Valvular heart diseases manifest as murmurs arising from turbulent blood flow, exhibiting distinct variations in intensity, timing, and frequency [5,6]. Despite being non-invasive and cost-effective, classical auscultation is inherently subjective, relying heavily on clinician expertise, and is further compromised by environmental noise and physiological variability [7,8]. These limitations have driven growing interest in automated computational methods for PCG analysis.

Deep learning has significantly advanced automated biomedical signal analysis, with CNNs and hybrid architectures successfully applied to murmur detection, heart sound classification, and cardiac pathology identification from raw signals or time-frequency representations [9–13]. Alkhodari and Fraiwan [18] proposed a CNN-BiLSTM architecture achieving an AUC of 0.998 for PCG-based VHD detection, while Jamil and Roy [19] introduced a Vision Transformer (ViT) framework achieving 99.90% accuracy and 99.95% F1-score on time-frequency PCG representations. However, key challenges persist: PCG signals are non-stationary and noise-prone [14]; datasets frequently exhibit class imbalance [15]; and clinical adoption demands model interpretability [16]. Furthermore, most existing studies address binary or single-pathology classification and adopt a single representation strategy without systematic comparison across signal modalities [17].

To address these gaps, this study proposes a dual-stream diagnostic pipeline integrating interpretable handcrafted feature analysis with deep spectrogram-based representation learning, evaluated on the clinically realistic BMD-HS dataset. A clinically grounded explainability framework is incorporated to identify physiologically meaningful spectral regions associated with pathological murmurs, supporting future deployment in AI-assisted digital stethoscope systems for affordable cardiac screening in resource-constrained settings.

1.1 Contributions

The main contributions of this work are:

– A multi-label framework is proposed for automated screening of co-occurring valvular heart diseases using the publicly available BMD-HS dataset under clinically realistic noise and recording variability.

- A systematic comparative analysis is conducted across handcrafted acoustic descriptors, raw waveform-based 1D-CNNs, and Mel-spectrogram-based 2D-CNNs within a unified experimental framework.
- A spectrogram-driven 2D-CNN achieves an AUC of 0.92 and a Normal-class F1-score of 0.94, demonstrating strong discrimination between pathological and healthy heart sounds.
- A clinically interpretable explainability framework combining feature-importance analysis and Grad-CAM-based spectral localization confirms model predictions are driven by medically relevant murmur frequency regions.
- The framework supports future deployment in AI-assisted digital stethoscope systems for low-cost cardiac screening in resource-constrained environments.

2 Data Description and Exploratory Data Analysis (EDA)

2.1 Dataset Overview

The BMD-HS (Bangladesh Medical College – Heart Sounds) dataset used for this study, comprising of phonocardiogram (PCG) recordings collected from the National Institute of Cardiovascular Diseases, Dhaka. It comprised of 108 patients, each contributing up to 8 recordings from diverse auscultation points (Mitral, Tricuspid, Pulmonary, Aortic) and postures (Supine, Sitting), adding almost 872 audio files. All recordings are sampled at 4000 Hz, enough to capture vital high-frequency murmurs (up to 600–700 Hz) and heart sounds (20–200 Hz). Digital stethoscopes is used to capture recordings in a real-world clinical environment, which results in, speech artifacts, ambient noise, and sensor friction.

2.2 Multi-label and Class Imbalance Characteristics

A key feature of the dataset is its multi-label nature: patients may exhibit multiple co-occurring valvular pathologies e.g., mitral stenosis with aortic regurgitation. Figure 1 represent the class distribution of BMD-HS dataset.

- Pathological Cases (Sick): Roughly balanced counts among Aortic Stenosis (37), Aortic Regurgitation (43), Mitral Regurgitation (38), and Mitral Stenosis (38).
- Normal Cases (Healthy): Only 21 subjects labeled strictly as Normal (20% of the dataset).
- Implication: The scarcity of normal samples increases the risk of over-diagnosis. To address this, data augmentation techniques (pitch shifting, noise injection) are applied to the Normal class during training.

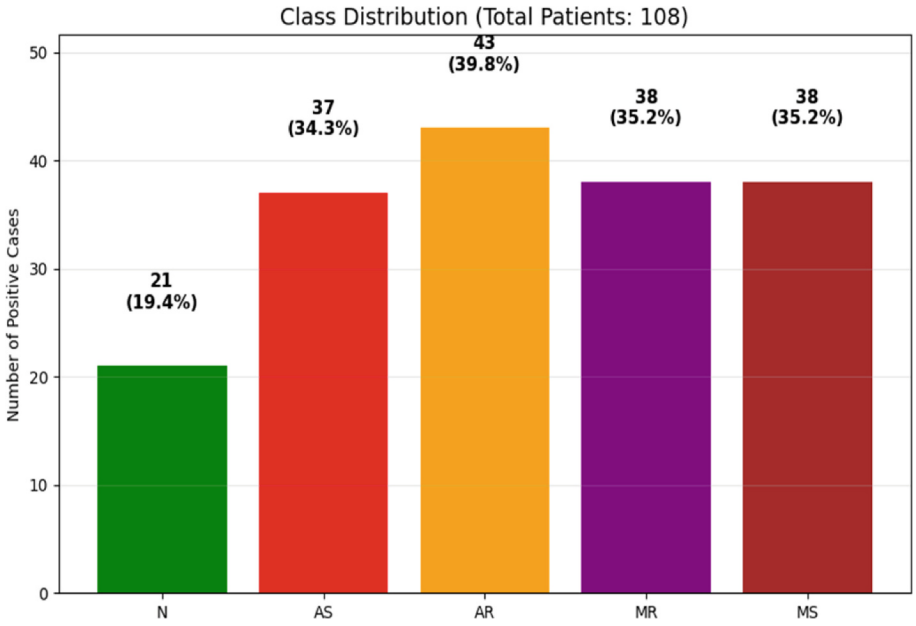

Fig. 1. The demonstration of class distribution of heart sound categories in the BMD-HS dataset comprising 108 patients.

2.3 Signal Characteristics

PCG signals are vary across patients and non-stationary. As represented in Fig. 2, the time-domain demonstration shows deviations in temporal structure and amplitude across cardiac cycles:

- Time: Evenly 20 s per recording.
- Amplitude: Active range differs due to stethoscope pressure and patient analysis. Signals are Z-score normalized to $[-1, 1]$.
- Noise Profile: Graphical examination exposes high-frequency sensor hiss and respiratory noise .

2.4 Spectral Analysis (Frequency Domain)

To examine spectral variances, time–frequency investigation was implemented using spectrogram representations and FFT, as presented in the right half of Fig. 2.

- Normal Signals: Energy determined in 20–200 Hz; S1 and S2 peaks are distinct and sharp, with clear quiet periods.
- Pathological Signals: Quiet intermissions interchanged by continuous noise (murmurs); substantial energy leakage in 200–600 Hz, emphasizing high-frequency modules as discriminative structures.

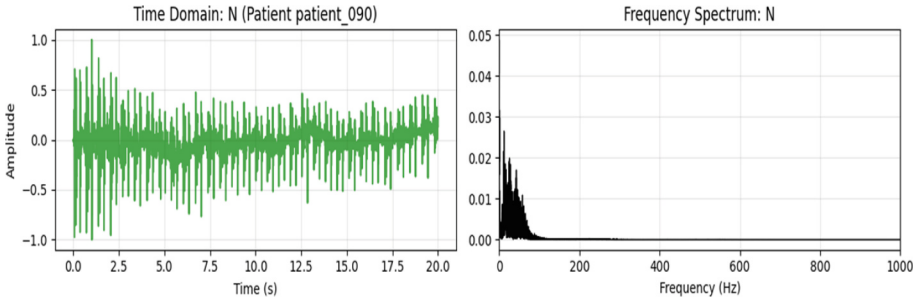

Fig. 2. Typical time-domain waveforms and their respective frequency-domain representations of PCG signals.

3 Methodology

The raw phonocardiogram (PCG) transforms recordings into automated diagnostic predictions through a proposed structured pipeline consisting of three main stages: signal preprocessing, data augmentation, and feature extraction and finally modeling. The suggested method enables the system to handle noisy real-world recordings while extracting significant acoustic features connected with cardiac irregularities. A advanced summary of the offered pipeline is demonstrated in Fig. 3.

3.1 Signal Preprocessing

The acquired PCG recordings in clinical environments frequently comprise of noise from ambient sounds, respiration, sensor friction and patient movement. A uniform preprocessing pipeline was applied to all recordings to enhance signal excellence and ensure robust feature extraction.

Resampling: At the native sampling rate of 4000 Hz all recordings were preprocessed. Sustaining this rate preserves significant high-frequency modules accompanying with pathological murmurs though escaping excessive loss of spectral evidence.

Bandpass Filtering: A second-order Butterworth bandpass filter with cutoff frequencies of 25 Hz and 400 Hz was applied. The ultimate heart sounds (S1 and S2) are typically focused below 150 Hz, while pathological murmurs may spread to higher frequencies. Frequencies below 25 Hz resemble mainly to standard wander produced by respiration, while frequencies above 400 Hz are mainly controlled by sensor noise.

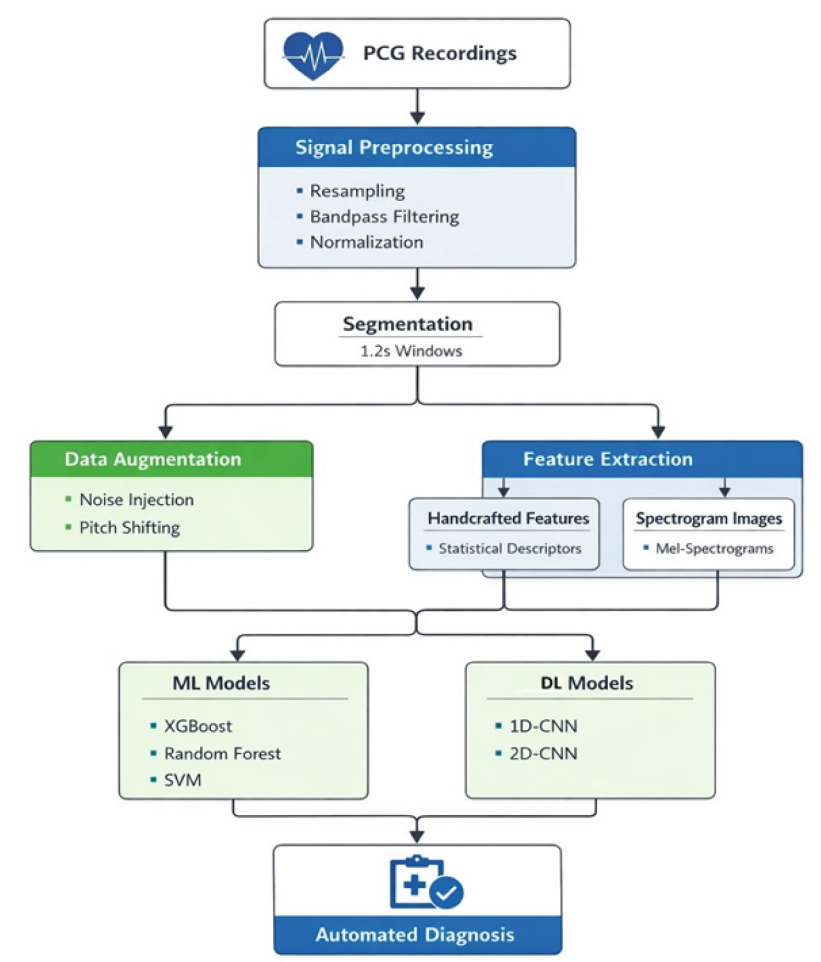

Fig. 3. The demonstration of workflow of proposed PCG-based automated diagnosis framework.

Amplitude Normalization: The recording volume variations caused by different stethoscope placements and patient physiology, signals using Min–Max scaling to normalize to the range [-1,1]. This guarantees that methods learn from waveform morphology rather than amplitude deviations.

Segmentation: The 20-second record was cut into 1.2-second pieces with an overlap of 50%. Medical scientists imagine a typical heart cycle at a rate of 60–100 beats per minute to be from 0.6 to 1.0 s. In this way, the window size chosen ensures that each segment of data has at least one full heartbeat.

3.2 Data Augmentation

There was a significant class imbalance in dataset, with only approximately 20% of normal subjects. To resolve this problem, data augmentation was applied specifically to the normal category during training. Two augmentation methods were used:

- **Noise Injection:** To simulate the low quality recordings Gaussian noise ($\mu = 0$, $\sigma = 0.005$) was added.
- **Pitch Shifting:** To emulate anatomical variations in heart size and acoustic propagation, signals were shifted by ± 2 semitones .

The applied data agumentation techniques increase the significant number of normal training samples. It also preserves the underlying acoustic characteristics of healthy heart sounds.

3.3 Feature Extraction

To leverage both the interpretability of classical machine learning and the pattern-recognition capabilities of deep learning, a dual feature extraction strategy was adopted. This approach extracts handcrafted statistical features for conventional machine learning models and deep visual features for convolutional neural networks.

Handcrafted Features: Handcrafted features capture mathematically defined properties of PCG signals and provide interpretable insights into acoustic characteristics associated with valvular abnormalities, as illustrated in Fig. 4.

Time-Domain Features.

- **Statistical Moments:** Mean, variance, skewness, and kurtosis describe the probability distribution of signal amplitudes.
- **Energy Metrics:** Root Mean Square (RMS) energy represents signal loudness.
- **Zero-Crossing Rate (ZCR):** Measures the frequency of sign changes in the signal and acts as a proxy for frequency content.
- **Crest Factor:** Ratio between peak amplitude and RMS value, indicating the impulsiveness of the waveform.

Frequency-Domain Features.

- **Mel-Frequency Cepstral Coefficients (MFCCs):** The first 13 coefficients were extracted to represent short-term spectral characteristics of the signal.
- **Spectral Centroid:** Indicates the center of mass of the spectrum and correlates with perceived brightness of the sound.
- **Spectral Entropy:** Measures the disorder of the spectral energy distribution.

Time–Frequency Features (Wavelet Analysis): Since PCG signals are non-stationary, wavelet-based analysis was used to capture localized time–frequency characteristics.

- A Discrete Wavelet Transform (DWT) using the Daubechies 4 (db4) mother wavelet was applied.
- The signal was decomposed into four levels.
- Energy and entropy were computed for both approximation and detail coefficients.

Feature Selection: To avoid the curse of dimensionality and remove redundant information, Principal Component Analysis (PCA) was applied to the aggregated feature set while retaining 99% of the cumulative variance.

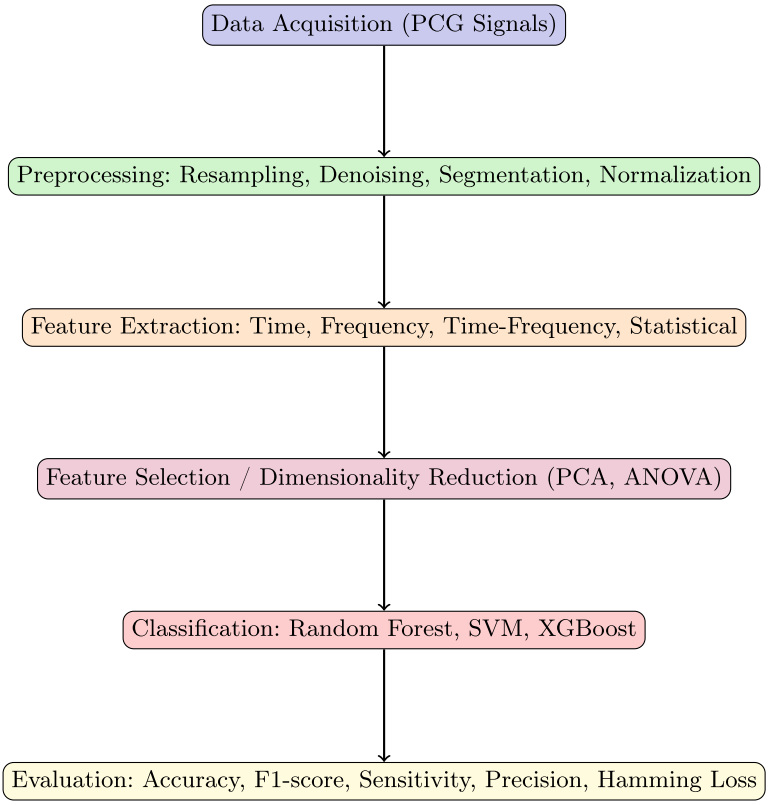

Fig. 4. A specific feature extraction pipeline carefully developed for classifying PCG signals.

Deep Visual Features. For convolutional neural network (CNN) models, the PCG classification problem was reformulated as a computer vision task by converting one-dimensional audio signals into two-dimensional time–frequency representations.

3.4 Mel-Spectrogram Generation Parameters

- Sampling rate: 4000 Hz
- FFT window size: 256 samples (∼64 ms)
- Hop length: 64 samples (∼16 ms overlap)
- Number of Mel bands: 64

The resulting power spectrogram was converted to the decibel (dB) scale to reflect the logarithmic sensitivity of human hearing. Each 1.2-second segment was therefore transformed into a 64 × 75 pixel Mel-spectrogram image, enabling CNN models to learn discriminative temporal–spectral patterns associated with cardiac abnormalities (Fig. 5).

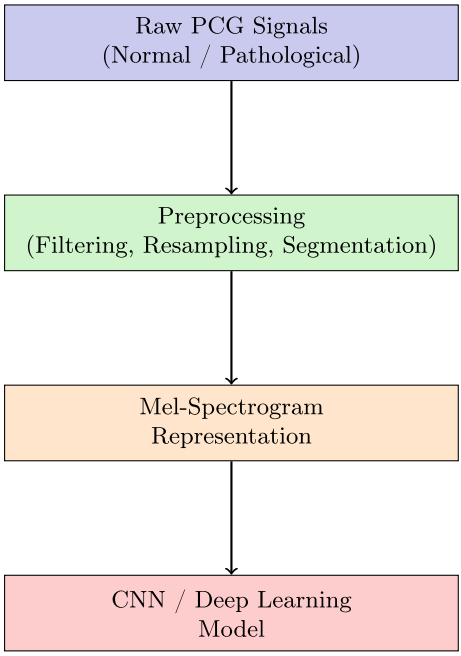

Fig. 5. The Pipeline represents the mel-Spectrogram for normal and pathological PCG Segments.

3.5 Modeling Strategy

Two complementary modeling strategies were employed.

- **Classical Machine Learning:** Handcrafted feature vectors were used to train interpretable models including XGBoost, Random Forest, and Support Vector Machines (SVM).
- **Deep Learning Models:** Spectrogram images were processed using 1D-CNN (raw waveform analysis) and 2D-CNN architectures operating on Mel-spectrogram representations.

This dual-stream framework combines the interpretability of statistical features with the representation learning capability of deep neural networks, enabling robust and clinically meaningful predictions.

4 Model Architectures and Comparative Analysis

To comprehensively evaluate different learning paradigms for heart sound classification, a diverse set of models was implemented. These models range from classical machine learning algorithms to deep neural networks and differ primarily in the type of input representation they process. Specifically, the models are grouped into three categories:

- **Feature-Based Learning (Tabular Data)** – models trained on handcrafted statistical, spectral, and wavelet features.
- **Time-Domain Representation Learning** – deep networks trained directly on raw PCG waveforms.
- **Time–Frequency Representation Learning** – convolutional models trained on Mel-spectrogram images.

This multi-model strategy enables a systematic comparison between interpretable feature-driven methods and deep representation learning approaches.

4.1 Feature-Based Learning (Tabular Data)

The first category of models utilizes the handcrafted feature vector derived from statistical, spectral, and wavelet features. After extraction, the feature space was reduced using Principal Component Analysis (PCA) to remove redundancy while retaining 99% of the variance. These models assume that the engineered features sufficiently capture the acoustic signatures associated with cardiac abnormalities.

4.2 Classical Machine Learning Classifiers:

Three widely used machine learning algorithms were evaluated as baseline models.

Random Forest (RF): Random Forest is an ensemble learning method consisting of multiple decision trees trained using the bagging strategy. In this study, a forest of **200 decision trees** was used. Random Forest was selected due to its robustness to noise and its ability to model complex non-linear interactions between features, such as relationships between wavelet energy and zero-crossing rate.

Support Vector Machine (SVM): Support Vector Machines were implemented with a **Radial Basis Function (RBF) kernel**. For the identification of optimal hyperplane that maximizes the margin between classes SVMs serves as high-dimensional feature spaces. This makes them perfect for discriminating subtle differences between pathological PCG and normal signals.

XGBoost (Extreme Gradient Boosting): To address the class imbalance in the dataset, the `scale_pos_weight` parameter was used to penalize misclassification of minority class samples. XGBoost was used for this purpose which is a gradient boosting algorithm that builds decision trees sequentially, where each new tree attempts to correct the errors of the previous ones.

Multi-layer Perceptron (MLP): A fully connected neural network was trained on the tabular feature vectors to bridge the gap between classical machine learning and deep learning.

- Input layer conforming to the PCA-reduced feature vector
- First hidden layer: 128 neurons with ReLU activation
- Second hidden layer: 64 neurons with ReLU activation
- Dropout layers (rate = 0.3)
- Output layer with sigmoid activation for classification

This architecture allows the model to pick up non-linear relationships between structures while preserving comparatively low computational complexity.

4.3 Time-Domain Representation Learning (1D-CNN)

A 1D Convolutional Neural Network (1D-CNN) was developed to learn discriminative features directly from the raw PCG waveform. Each input sample comprised of a raw audio segment with a length of 4800 samples, corresponding to $1.2\,s$ of audio recorded at a sampling rate of $4000\,Hz$. The proposed ID-CNN was developed to learn hierarchical temporal demonstrations from the PCG signals.

Initial convolutional layers operate with larger kernels (size = 9) to the low-frequency temporal patterns, for example, the rhythm and periodicity related to the S1 and S2 heart sounds. Whereas for the detection of high-frequency transient patterns deeper convolutional layers employ smaller kernels (size = 3). A GlobalAveragePooling1D layer was applied before the final classification stage instead of traditional fully connected layers to reduce the model complexity and improve generalization. This procedure breakdown the temporal dimension by computing the average activation of each feature map, by this means meaningfully dropping the number of trainable parameters and decreasing the risk of overfitting, which is particularly important when working with moderate-sized biomedical datasets.

4.4 Time–Frequency Representation Learning (2D-CNN)

The Mel-spectrogram images used as a input for convolutional neural network for classification and this is the third modeling approach. In this approach, each PCG segment is converted into a 64 × 75 Mel-spectrogram image , which demonstrats the time–frequency energy distribution of the heart sound signal. A VGG-19 CNN based architecture was implemented to process these spectrogram representations. The model comprised of three blocks, Conv2D layer, followed by Batch Normalization, a ReLU activation function, and a MaxPooling2D layer. These arranged convolutional blocks gradually extract higher-level spectral features from the Mel-spectrogram images. The 2D convolution filters operate across both the time axis (horizontal dimension) and the frequency axis (vertical dimension) unlike 1D-CNN models that analyze signals purely in the temporal domain. This allows the network to learn spatiotemporal features within the spectrograms, permitting it to recognize visually diverse acoustic features such as broadband noise areas related with murmurs and transient spectral peaks equivalent to valve clicks.

4.5 Training Configuration

A consistent training configuration was adopted to ensure a fair comparison across all neural network models (MLP, 1D-CNN, and 2D-CNN). A Binary Cross-Entropy loss function and the Adam optimizer with a learning rate of 0.001 used for all model trainings. A ReduceLROnPlateau scheduler was employed to reduce the learning rate when the validation loss stopped improving, whereas early stopping was employed to dismiss training if the validation loss did not improve for 10 successive epochs. This standardized training protocol guaranteed that performance differences primarily returned architectural proficiencies relatively than optimization setting variations (Table 1).

5 Results and Discussion

The quantitative evaluation of the proposed models for heart sound classification is presentsed in this section. The 80/20 train-test split was performed at the

Table 1. Training Parameters

Parameter	Value/Description
Framework	Python, TensorFlow/Keras
Maximum Epochs	100
Batch Size	32
Early Stopping Patience	10 epochs (monitored on validation loss)
Train–Test Split	80/20 stratified split
Validation Set	Subset of training partition
Intermediate Activation	ReLU
Output Activation	Sigmoid
Loss Function	Binary Crossentropy
Optimizer	Adam
Learning Rate	0.001
LR Scheduling	ReduceLROnPlateau
Classification Type	Multi-label

patient level, ensuring that all segments derived from a given patient appeared exclusively in either the training or test set. This strict partitioning eliminates the risk of data leakage and ensures that the model is evaluated on entirely unseen patients rather than unseen segments from familiar subjects. The Model was evaluated on the held-out test set using metrics appropriate for imbalanced multi-label classification. In addition to classical machine learning baselines, the evaluation framework was designed to compare handcrafted feature learning, raw waveform representation learning, and spectrogram-based deep learning strategies in order to comprehensively assess the effectiveness of different PCG signal representations for valvular heart disease screening.

5.1 Evaluation Metrics

Traditional accuracy alone is insufficient for evaluating multi-label medical classification systems. A model predicting only the majority class may achieve high accuracy while failing to detect diseases. Therefore, multiple complementary metrics were used.

Exact Match Accuracy: measures the percentage of samples for which the predicted label set exactly matches the ground truth:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i)$$

Since this metric is strict for multi-label problems, we primarily rely on the

Macro F1-Score: averages the F1-score across all classes:

$$Macro\ F1 = \frac{1}{L}\sum_{l=1}^{L}\frac{2P_l R_l}{P_l + R_l}$$

In addition, two clinically relevant metrics were considered:

Recall (Sensitivity): measures the ability to detect pathological cases.

Precision: measures the reliability of positive predictions.

Hamming Loss: was used to measure the fraction of incorrectly predicted labels:

$$HammingLoss = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{l=1}^{L} xor(y_{i,l}, \hat{y}_{i,l})$$

Lower Hamming Loss indicates better multi-label prediction performance.

5.2 Performance of Feature-Based Models

Table 2 summarizes the performance of models trained on the handcrafted feature vector consisting of statistical, spectral, and wavelet features. Classical machine learning models achieved relatively low exact-match accuracy, indicating difficulty in capturing complex dependencies between multiple cardiac conditions. Among them, XGBoost produced the strongest baseline performance with a Macro F1-score of 0.61, demonstrating its ability to detect pathological patterns in the feature space.

Table 2. Performance of feature-based models

Model	Accuracy	Macro F1	Hamming Loss
Random Forest	24.8%	0.49	0.22
SVM (RBF)	28.5%	0.52	0.20
XGBoost	35.2%	0.61	0.18
MLP	89.0%*	0.68	0.12

The Multi-Layer Perceptron (MLP) significantly improved performance, achieving a Macro F1-score of 0.68. This suggests that relationships among spectral and wavelet features are highly nonlinear and benefit from neural network representations.

5.3 Performance of Representation Learning Models

Deep learning models trained directly on signal representations demonstrated further improvements in classification performance. Table 3 compares the results of the 1D-CNN (time-domain) and 2D-CNN (spectrogram-based) architectures.

Table 3. Deep learning model comparison

Metric	1D-CNN	2D-CNN
AUC	0.80	0.92
Macro F1	0.59	0.76
Precision	0.58	0.76
Recall	0.77	0.81
Normal Class F1	0.73	0.94

The 1D-CNN demonstrated high recall for pathological classes, indicating strong sensitivity for detecting abnormal heart sounds. This marks it appropriate as a screening model. Though, the model demonstrated comparatively low precision due to difficulty differentiating noisy common signals from murmurs. Comparatively, the 2D-CNN trained on Mel-spectrogram illustrations attained the best overall presentation with an AUC of 0.92. The model considerably improved classification of normal cases, achieving a normal class F1-score of 0.94. These results propose that spectrogram demonstrations provide richer time–frequency information, which enables the CNN to learn acoustic features connected with cardiac murmurs more efficiently. Experimental observations showed that smaller FFT windows weakened the spectral discrimination of murmurs, while larger windows introduced temporal smoothing that hurt the localization of transient acoustic patterns.

5.4 Comparison with State-of-the-Art

Table 4. Comparison with existing state-of-the-art methods for valvular heart disease detection using PCG signals

Study	Method	Dataset	AUC	F1/Accuracy
Alkhodari & Fraiwan (2021)	CNN-BiLSTM	Private clinical	0.998	99.32% Acc
Jamil & Roy (2023)	Vision Transformer (ViT)	Private PCG	__	99.95 F1
Proposed	**2D-CNN (Mel-spectrogram)**	**BMD-HS**	**0.92**	**0.76 F1**

To contextualize the performance of the proposed framework, Table 4 presents a comparison with closely related state-of-the-art studies on PCG-based valvular heart disease detection. Alkhodari and Fraiwan [18] proposed

a CNN-BiLSTM architecture achieving an AUC of 0.998 on a private clinical dataset. Jamil and Roy [19] employed a Vision Transformer (ViT) with an accuracy of 99.90% and F1-score of 99.95% on a private PCG dataset. While these results appear higher, it is important to note that both studies operate on private datasets under binary or limited-class settings, whereas the proposed framework addresses the more challenging and clinically realistic problem of multi-label classification on the publicly available BMD-HS dataset.

5.5 Explainability Insights

Explainability techniques were applied to both feature-based and deep learning models to interpret model predictions. XGBoost model identified spectral centroid and wavelet energy as the most discriminative patterns. These conclusions align with clinical knowledge, as pathological murmurs naturally display amplified high-frequency energy due to turbulent blood flow. Additionally, Grad-CAM visualizations of the 2D-CNN discovered that the model emphases mainly on spectral regions between 200–600 Hz during systolic and diastolic intervals. This authorizes that the network learns medically appropriate acoustic features rather than background noise. It improves confidence in the model's diagnostic predictions. While Grad-CAM visualizations offered meaningful insight into the spectral regions driving model predictions, quantitative clinical validation of explainability remains an important area for future work. In this study, the highlighted regions were interpreted against established clinical characteristics of valvular murmurs from the cardiology literature, particularly the concentration of abnormal acoustic energy in higher-frequency bands. Future work will involve expert cardiologist evaluation and quantitative localization analysis to further validate the clinical reliability of the explainability framework.

6 Conclusion and Future Work

This research work proposed an end-to-end deep learning pipeline for the automated multi-label classification of valvular heart diseases using PCG recordings. The proposed system addresses key challenges in clinical datasets, including signal noise, co-occurring pathologies, and class imbalance by integrating handcrafted feature analysis with deep representation learning. Experimental results proves that the 2D-CNN architecture trained on Mel-Spectrogram demonstration gains the highest predictive performance, with an AUC of 0.92 and a Normal class F1-score of 0.94. The model effectively mitigates the over-diagnosis issues observed in 1D and classical feature-based models by treating heart sounds as visual textures.

– Clinical Significance: The model shows high clinical significance by maintaining high sensitivity for critical conditions such as Aortic Regurgitation (Recall $> 80\%$). It proves that its potential use as a low-cost, non-invasive screening tool in resource-constrained healthcare settings.

– Trustworthiness: Th model focuses on medically relevant spectral features with the help of post-hoc explainable AI analyses, ensuring that predictions are guided by physiologically meaningful patterns rather than dataset artifacts.

6.1 Future Work

Future research will focus on improving the accessibility, utility, and generalization of the proposed framework. One encouraging direction is to incorporate the 1D-CNN, which displays high sensitivity, with the 2D-CNN, which offers high specificity, into a weighted ensemble model. This will potentially optimize the balance between recall and precision for clinical deployment. Furthermore, model optimization and quantization for embedded platforms, such as Raspberry Pi devices or smartphone, could enable a standalone AI-powered stethoscope for point-of-care cardiac screening. Finally, framework evaluation on independent datasets, such as PhysioNet, will be essential to assess its generalizability across different recording devices and diverse patient populations. It ensures robust real-world performance. Overall, the proposed method provides a interpretable, robust, and low-cost solution for automated cardiac screening, which represents a significant step toward accessible AI-driven healthcare diagnostics.

References

1. Ayappasamy, K., et al.: Detection of valvular heart diseases from PCG signals using machine and deep learning models–a review. IEEE Access (2025)
2. Pavithra, G.R., et al.: Enhanced cardiovascular disease detection from phonocardiogram signals using deep learning & wavelet-based denoising. In: 2024 4th Asian Conference on Innovation in Technology (ASIANCON). IEEE (2024)
3. Muhammad Rafli, R., et al.: Enhanced identification of valvular heart diseases through selective phonocardiogram features driven by convolutional neural networks (SFD-CNN). Jurnal Nasional Teknik Elektro, 20–35 (2024)
4. Anjan, G., et al.: Automated system for the detection of heart anomalies using Phonocardiograms: a systematic review. IEEE Access **12**, 138399–138428 (2024)
5. Hend Khalid, A., et al.: Precision diagnosis: an automated method for detecting congenital heart diseases in children from phonocardiogram signals employing deep neural network. IEEE Access **12**, 76053–76064 (2024)
6. Zihan, J., et al.: Automated valvular heart disease detection using heart sound with a deep learning algorithm. IJC Heart Vasculature **51**, 101368 (2024)
7. Bollapalli, A., Panini Challa, N.: Explainable attention-based deep learning for classification and interpretation of heart murmurs using phonocardiograms. Sci. Reports **15**(1), 37991 (2025)
8. Subham Kumar, P., Mohapatra, A., Patra, S.: X-CBNet: an explainable effective deep learning framework based on spectrograms for predicting valvular disorder using PCG signals. J. Transform. Technol. Sustain. Dev. **9**(1), 18 (2025)
9. Popal Khan, P., et al.: Enhancing cardiac health diagnoses through machine learning analysis of phonocardiograms (PCG). J. Data Sci. Intell. Syst. **3**(4), 312–323 (2025)

10. Abdul, J., et al.: Automated detection of pediatric congenital heart disease from phonocardiograms using deep and handcrafted feature fusion. Comput. Biol. Med. **197**, 110993 (2025)
11. Hussain, S.S., Ashfaq, M., Khan, M.S., Anwar, S.: Deep learning based phonocardiogram signals analysis for cardiovascular abnormalities detection. In: 2023 International Conference on Robotics and Automation in Industry (ICRAI), Peshawar, Pakistan, pp. 1-6 (2023). https://doi.org/10.1109/ICRAI57502.2023.10089537
12. Ihtisham Ul, H., et al.: Enhancing pediatric congenital heart disease detection using customized 1D CNN algorithm and phonocardiogram signals. Heliyon, **11**(3) (2025)
13. Hemalatha, G., et al.: Precision diagnosis: an automated method for detecting congenital heart disease in children from phonocardiogram signals employing deep neural network. In: Recent Trends in Data Analytics and Computing, pp. 223–228. CRC Press (2024)
14. Orozco-Reyes, L., Alonso-Arévalo, M.A., García-Canseco, E., Ibarra-Hernández, R.F., Conte-Galván, R.: A deep-learning approach to heart sound classification based on combined time-frequency representations. Technologies **13**, 147 (2025). https://doi.org/10.3390/technologies13040147
15. Sukant, B., Saha Misra, I., Naveed Siddiqui, K.: Phonocardiogram signal based better prediction of aortic and mitral stenosis heart diseases using machine learning. In: 2025 International Conference on Ambient Intelligence in Health Care (ICAIHC). IEEE (2025)
16. Sukant, B., Saha Misra, I., Naveed Siddiqui, K.: Detection of aortic and mitral stenosis valvular heart diseases using machine learning on PCG signals. In: 2024 IEEE Calcutta Conference (CALCON). IEEE (2024)
17. Soh-Yeon, L., et al.: Deep learning-based evaluation of the severity of mitral regurgitation in canine myxomatous mitral valve disease patients using digital stethoscope recordings. BMC Veterinary Res. **21**(1), 326 (2025)
18. Alkhodari, M., Fraiwan, L.: Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. Soft. Comput. **25**, 7017–7030 (2021)
19. Jamil, S., Roy, A.M.: An efficient and robust Phonocardiography (PCG)-based Valvular Heart Diseases (VHD) detection framework using Vision Transformer (ViT). Biomedical Signal Processing and Control (2023)