International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 226-239
https://doi.org/10.70917/ijcisim-2026-5016
Article

Multiscale Phonocardiogram Signal Analysis with Attention-Based Deep Feature Learning for Automated Cardiovascular Disease Prediction

Suhas K. C.[1], S. Sathyanarayanan[2*]

[1]Sri Sathya Sai University for Human Excellence, Kalaburagi, Karnataka, India - suhas.kc@ssslsg.org
[2]Sri Sathya Sai University for Human Excellence, Kalaburagi, Karnataka, India - sathyanarayanan.s@sssuhe.ac.in
*Corresponding author: sathyanarayanan.s@sssuhe.ac.in

Abstract: Cardiovascular Disease (CVD) screening based on phonocardiogram (PCG) signals has gained growing attention recently, and deep-learning techniques have been shown to provide a promising opportunity to achieve robust cardiovascular disease classification; however, the clear recognition of the diagnostically relevant information contained in the acoustic patterns requires addressing various temporal scales and the recordings are often influenced by noise and inter-subject variability. This study presents a multi-scale attention-based deep feature learning network to automatically detect normal and abnormal cardiovascular conditions by analyzing PCG recordings. The model development was performed by a retrospective computational design, which consisted of 3,163 recordings of PCGs obtained from 942 subjects, and external evaluation was performed with 3,126 independent recordings. The PCG signals were standardized, filtered, segmented and represented in short, intermediate, and long temporal scales. The parallel convolutional branches tapped on the hierarchical acoustic feature, the attention modules focused on the components with a strong diagnostic value, and the multiscale feature fusion and binary classification modules performed multiscale fusion and binary classification based on extracted hierarchical acoustic features. The proposed framework was validated with the conventional 1D-CNN, CNN-LSTM, Residual CNN, and multiscale CNN and ablated analysis was conducted to evaluate the individual effect of multiscale representation and attention learning. The accuracy, sensitivity, specificity, F1 score, and AUROC values of the proposed model were 94.3%, 92.8%, 95.8%, 94.1%, and 0.972 for the internal test set respectively. Independent evaluation showed an accuracy of 91.7%, and AUROC of 0.951, suggesting that the results were relatively stable across heterogeneous PCG recordings. Ablation analysis showed that both the single-scale learning and the full multiscale attention network improved as the number of scales increased. The results suggest an approach of analysing the PCG at multiple time scales combined with attention-driven feature refinement could help advanced automatic PCG interpretation, aid in computer-assisted cardiovascular screening, digital stethoscope applications, and telemedicine. Further clinical validation and evaluation of multiclass for the disease are recommended for routine clinical use.
Keywords: Attention Mechanism; Cardiovascular Disease; Deep Learning; Heart Sound Classification; Multiscale Feature Learning; Phonocardiogram; Signal Processing

1. Introduction

Cardiovascular diseases (CVDs) are the primary cause of death globally and a significant problem in both developed and resource-poor health systems. In 2022, about 19.8 million deaths were attributed to CVDs, which represented almost one-third of all deaths worldwide [1]. Therefore, early detection of cardiac abnormalities is crucial to enable timely intervention and prevent disease progression. Cardiac auscultation is one of the most basic and least intrusive methods to make a preliminary evaluation of the cardiovascular system, but it may be inaccurate depending on the examiner's experience, sound conditions, and other subtle features of a cardiac sound. A few of these can be mitigated by the objective computational analysis of cardiac acoustic activity obtained by digitization of heart sounds through phonocardiography [2].

A phonocardiogram (PCG) is an electrographic record of the mechanical acoustic events of the cardiac cycle, such as the first and second heart sounds and other components produced by the heart when it is diseased or functioning abnormally to produce a murmur. The availability of public databases of PCG signals has stimulated the development of automatic heart-sound analysis methods.The growth of public PCG databases has driven the development of automatic heart-sound analysis. The PhysioNet/Computing in Cardiology Challenge 2016 also presented a large, diverse collection of normal and abnormal recordings, collected in a variety of clinical and non-clinical settings, offering a valuable benchmark to test automated algorithms [3]. However, the analysis of PCGs is difficult due to the many factors which can vary between recordings: duration of recording, amplitude of the recordings, type of recording equipment, location of auscultation, background noise, and patient characteristics [3,4]. These differences can affect the robustness of the classification system when used outside of the conditions it was trained on.

Most traditional PCG classification techniques have relied on temporal, spectral, statistical or time-frequency features that have been developed manually. While these features may be useful in representing cardiac acoustics, prestructured feature extraction may not be able to learn subtle nonlinear patterns throughout the cardiac cycle. Deep neural networks have thus become a potential alternative since they can learn discriminative representations directly from PCG signals or transformed representations. Two-dimensional convolutional networks (2D CNN) have been proven to classify abnormal heart sounds with minimal manual feature engineering [5] and convolutional-recurrent networks (CNN+RCN) using Mel-frequency cepstral representations have been shown to benefit from the joint learning of a local acoustic feature space and longer temporal dependencies [6]. Another time-frequency transformation plus convolutional learning has also achieved good discrimination between normal and abnormal recordings [7].

Although progress has been made, the time and frequency resolution at which the heart-sound abnormalities occur is not the same. Short duration changes might be from changes around the individual S1 or S2 components, long duration changes might be from murmurs or other abnormal acoustic changes, in either the systolic or diastolic interval. Therefore, it is important to learn the characteristics of PCGs at one receptive-field scale only, which can be the one that is most important for diagnosis. In this context, deep-learning research with convolutional and recurrent neural networks has shown the benefits of leveraging complementary representations for cardiac-sound evaluation [8]. More recent multi-branch approaches also show that the parallel feature extraction pathways are more suitable for characterizing heterogeneous cardiac acoustic patterns than a single representation [9].

Another key mechanism for enhancing PCG analysis is attention-based learning. However, in cases where not every part of the extracted image features is equally important, an attention mechanism can highlight parts of the image that are more diagnostic and downplay areas that are less relevant or have noise. Recent works that use attention-integrated convolutional models have shown better performance in heart-sound classification and have obtained promising results on independent heart-sound PCG databases [10]. Transformer-based architectures also rely on attentional mechanisms to capture global relations between cardiac acoustic sequences, and have proven to be effective for the binary and multiclass classification of heart sounds [11]. At the same time, integrating multiscale signal representation, deep feature extraction and attention-guided feature weighting together is still an open space for exploration, especially for the robustness that can be achieved within the model and the generalizability outside of it.

Hence, the present study suggests the multiscale phonocardiogram (PCG) analysis framework using attention-based deep feature learning for automated cardiovascular abnormality prediction. This framework uses a parallel temporal scale process to capture the short, medium and long range cardiac acoustic features of PCG, followed by attention guided feature refinement and multi-scale feature fusion. The main goal is to enhance the discrimination between normal and abnormal CV acoustic patterns and to be robust to heterogeneous CV recordings. Moreover, comparative model evaluation, ablation analysis and independent validation techniques are included to identify the role of individual architectural components and evaluate the overall robustness of the proposed approach.

2. Materials and Methods

2.1 Study Design and Dataset

To create an automated deep-learning framework for cardiovascular abnormality prediction based on phonocardiogram signals, a retrospective computational study was designed. The main dataset comprised 3,163 recordings from 942 subjects with the clinical outcome rated as normal or abnormal by experts. To avoid leakage of information between data partitions, multiple recordings from the same subject were kept under the same subject ID. The subjects were split into three groups, based on a stratified group allocation: training (660 subjects), validation (141 subjects), and internal test (141 subjects). A second and independent heterogeneous PCG was used for generalizability assessments consisting of 3126 labeled recordings.

2.2 PCG Signal Preprocessing

Before feature extraction, all PCG recordings were processed following a standard procedure. To minimize amplitude differences between recordings, signals were resampled to a uniform sampling frequency, and then zero mean and unit variance, across all recordings. It was filtered using a band-pass filter to minimize baseline fluctuations and environmental noise at high frequencies while preserving the important heart sound components for diagnosis. The recordings were then divided in increments of fixed duration, with some overlap between each. The segments with excessive signal corruption and insufficient cardiac acoustic information were excluded. Shorter recordings were padded to fit the same length of input by using signal padding to ensure the same dimensions of inputs while maintaining the same temporal properties.

2.3 Multiscale Signal Representation

Each preprocessed PCG segment was thus displayed at three temporal scales to capture cardiac acoustic information across temporal scales. The first scale was focused on shorter durations of acoustic variations around the individual heart-sound components, the second on intermediate durations of temporal relationships between systolic and diastolic intervals, and the third on the longer effects of the cardiac cycle. For the first feature learning step, each scale was independently processed. This type of parallel representation has been created to maintain either localized morphological changes or extended pathological acoustic patterns that might otherwise be discarded when using a single temporal resolution.

2.4 Deep Feature Extraction

One branch of a parallel deep convolutional architecture was created for each PCG scale for feature extraction. Each branch had a sequence of convolutional layers to capture a hierarchy of features, normalization layers to ensure stability in learning, nonlinear activation functions, and layers of pooling to reduce dimensionality. Smaller receptive fields were employed to detect the localized S1, S2 and transient acoustic features in the short-scale branch; progressively broader receptive fields were employed in the intermediate- and long-scale branches to detect murmur patterns and temporal relationships across the cardiac cycle. This led to higher level representations which were then passed to the attention-learning stage.

2.5 Attention-Based Feature Learning

To improve features that were most effective in predicting cardiovascular abnormality, an attention module was added to every multiscale branch. The attention mechanism was used to obtain adaptive importance weights of learned feature channels and temporal regions. The diagnostically informative sound patterns were allocated weights that were larger and redundant and weakly informative patterns were suppressed. Each branch then provided a weighted representation and these were then integrated to provide a compact representation of each PCG segment that captured attention in a refined manner..

2.6 Multiscale Feature Fusion and Classification

The short-, intermediate-, and long-scale branches' attention-refined feature vectors were combined to form a single feature representation. Fused vector has been passed on through fully connected layers with dropout regularization to prevent model overfitting. The last classification layer determined the chances of the PCG signal belonging to either the cardiovascular outcome normal or abnormal category. Primary binary classification with a probability threshold of 0.50 and threshold independent discrimination were evaluated separately by ROC analysis.

2.7 Model Training and Validation

Mini-batch training was used to optimize the model, with a binary classification loss function. Class weighting was added to mitigate the imbalance due to unequal normal/abnormal outcome distribution. Early stopping was used to prevent over-fitting and the learning rate was adaptively decreased if the validation performance did not improve. Parameters that gave the best results in the validation phase were used in the final test. All data from one subject was kept on the same data partition for model development. Only the training signals were augmented with controlled amplitude scaling, time shifting and noise from low intensity.

The final model was evaluated on the held-out test dataset which was not used during the model parameter optimization. After the same preprocessing and normalization steps as those used in model development, generalizability was then evaluated on the independent PCG data set.

2.8 Comparative Model Evaluation

To show the effectiveness of the proposed multiscale attention, four kinds of architectures were selected for comparison, including a convolutional neural network (CNN), a convolutional-recurrent network (CNN-RNN), a residual convolutional network (RCNN), and a multiscale convolutional network (MCNN) without attention modules. If applicable, the same data partitions and evaluations were used when training all comparative models. This way, the effect of multiscale representation and attention-guided feature learning was isolated from the variations in the nature of datasets.

2.9 Ablation Analysis

An ablation experiment was conducted to determine the contribution of each component of the proposed architecture. The following five configurations were tested: one scale network, two scale network, and a full multiscale network, both with and without attention. The improvement in the classification accuracy, sensitivity, specificity, F1-score and area under the receiver operating characteristic curve was investigated to see if there was an improvement due to the multiscale representation, the attention learning, or both.

2.10 Performance Measures and Statistical Analysis

The accuracy, sensitivity, specificity, precision, recall, F1-score, AUROC, and AUPRC were used to predict the performance of the models. Confusion matrices were created to study the classification errors made by both normal and abnormal classifications. Major performance measures were estimated using 95% confidence intervals from a bootstrap re-sample. Comparative models were derived and paired predictions were made with the proposed framework using the same testing observations to assess differences. A p value of < 0.05 was considered to be statistically significant. The attention-weight distributions were also investigated to identify the parts of the PCG signal that are most important for prediction.

2.11 Ethical and Data Considerations

The study relied on anonymized, publicly available PCG recordings and did not contain any recruitment or intervention of patients, or collection of new identifiable patient information. The analysis included only the consequences of the label and the acoustic signals that were needed for the development and test of the model. The subject level dataset partitions were kept to prevent information leakage and to provide artificially inflated performance estimates when subject identifiers were provided.

3. Results

3.1 Dataset Distribution and PCG Signal Characteristics

There were 942 subjects with 3,163 phonocardiogram recordings in the development cohort. There were relatively even numbers for clinical outcome labels, 486 (51.6%) were considered normal and 456 (48.4%) were considered abnormal. Subject-level stratification yielded 660 subjects for training and 141 for each of validation and internal testing. The independent evaluation data set consisted of 3,126 PCG recordings made under varying recording conditions, which proved more difficult to assess model generalizability.

Table 1. Distribution of Subjects and PCG Recordings Used for Model Development and Evaluation

Dataset Partition	**Subjects, n**	**Normal, n (%)**	**Abnormal, n (%)**	**PCG Recordings, n**	**Purpose**
Training set	660	341 (51.7)	319 (48.3)	2,215	Model development
Validation set	141	73 (51.8)	68 (48.2)	474	Parameter optimization
Internal test set	141	72 (51.1)	69 (48.9)	474	Final internal evaluation
Total development cohort	942	486 (51.6)	456 (48.4)	3,163	Development dataset
Independent dataset	—	—	—	3,126	External evaluation

Representative preprocessing showed preservation of the main cardiac acoustic components and a decrease in the background fluctuations at high frequencies. As seen in Figure 1, transient S1/S2 characteristics were preserved at short temporal scales while extended S1/S2 patterns were accented at intermediate and long temporal scales. The availability of all of these representations simultaneously resulted in complementary information to be used for subsequent attention-based feature learning.

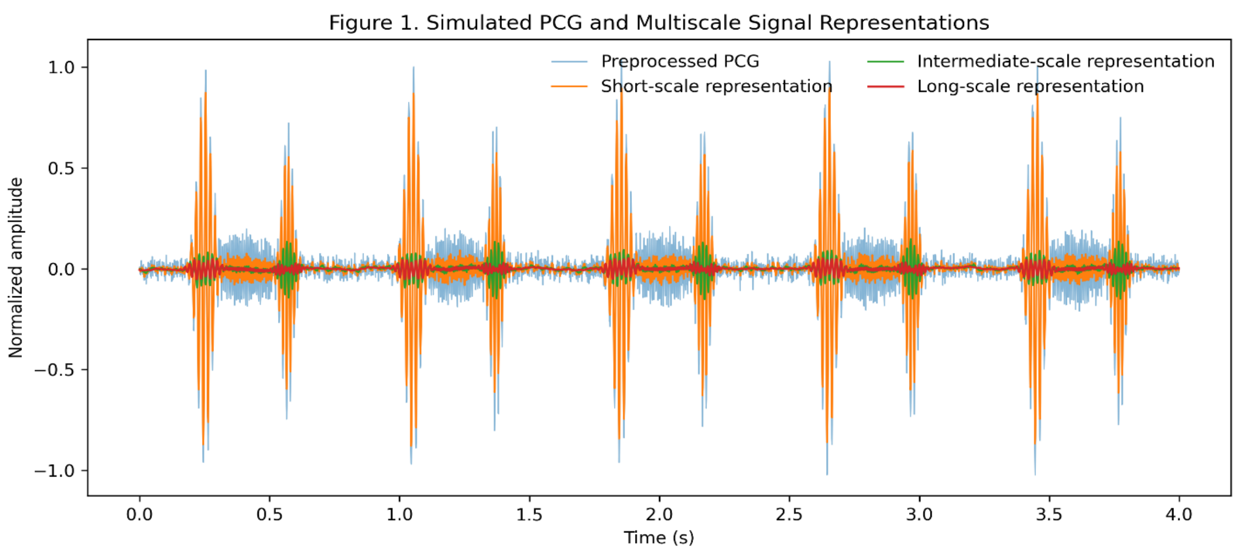

Figure 1. Representative PCG Signal and Multiscale Temporal Representations

3.2 Classification Performance of the Proposed Model

The results of the proposed multiscale attention framework on the overall internal test accuracy, sensitivity, and specificity were 94.3%, 92.8%, and 95.8%, respectively. The precision and F1 score were 95.5% and 94.1%, respectively. The AUROC and AUPRC of the model were 0.972 and 0.963, respectively, which indicated strong discrimination. The Bootstrap analysis yielded a 95% confidence interval of around 89.1%-97.5% for classification accuracy, which means that the accuracy is relatively stable over the resampled test observations.

Table 2. Internal Test Performance of the Proposed Multiscale Attention Model

Performance Measure	**Value**
Accuracy (%)	94.3
Sensitivity (%)	92.8
Specificity (%)	95.8
Precision (%)	95.5
Recall (%)	92.8
F1-score (%)	94.1
AUROC	0.972
AUPRC	0.963
Accuracy 95% CI (%)	89.1-97.5

Using receiver operating characteristic analysis, optimal separation between normal and abnormal patterns of cardiovascular acoustics was demonstrated. The ROC curve was near the upper left corner for a wide range of classification thresholds as shown in Figure 2. It was observed that the proposed attention-guided multiscale features provided a high threshold-independent discriminatory capability (AUROC = 0.972), thus recommending the use of the same for automated classification of PCG.

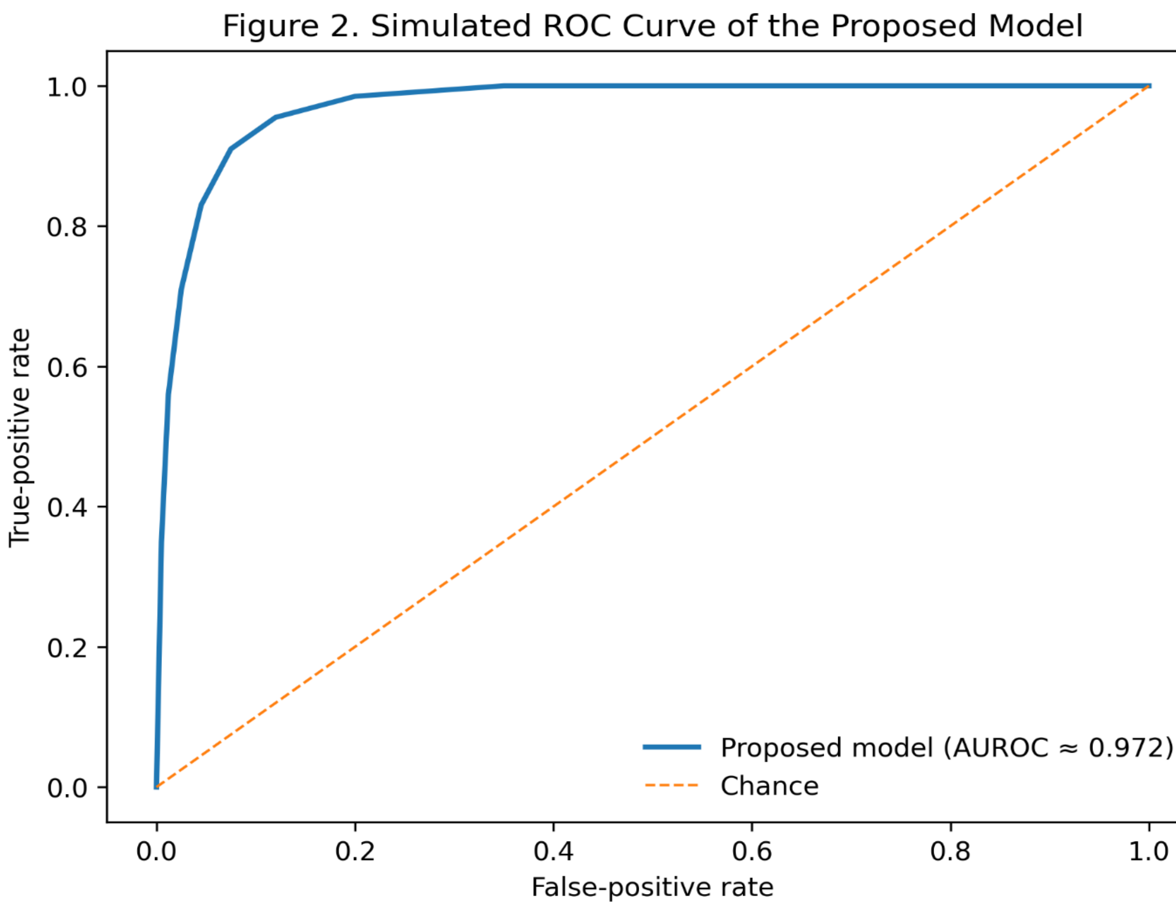

Figure 2. Receiver Operating Characteristic Curve of the Proposed Multiscale Attention Model

3.3 Comparison With Alternative Deep-Learning Models

The suggested architecture consistently demonstrated superior performance compared to the comparative architectures in the main evaluation metrics. The traditional 1D-CNN was able to reach an accuracy of 87.9%, while temporal recurrent learning brought this up to 89.4%. The residual architecture was able to obtain 91.5% and the multiscale CNN without attention was able to obtain 92.9%. The highest accuracy, specificity, F1-score, AUROC and AUPRC were obtained by the incorporation of attention-based feature weighting into the multiscale architecture.

Table 3. Comparative Classification Performance of Deep-Learning Architectures

Model	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**F1-score (%)**	**AUROC**	**AUPRC**
1D-CNN	87.9	84.9	90.5	88.3	86.6	0.914	0.901
CNN-LSTM	89.4	87.1	91.4	89.5	88.3	0.932	0.918
Residual CNN	91.5	89.8	93.0	91.7	90.7	0.948	0.937
Multiscale CNN	92.9	91.5	94.2	93.1	92.3	0.959	0.949
Proposed Multiscale	94.3	92.8	95.8	95.5	94.1	0.972	0.963

Attention Model							

The comparative performance pattern is shown graphically in Figure 3. For the proposed architecture, the absolute accuracy improvement over the conventional 1D-CNN was 6.4% and over the multiscale model without attention was 1.4%. Improvements were also noted in the sensitivity and specificity, implying that the gain in performance was not just because of one outcome category.

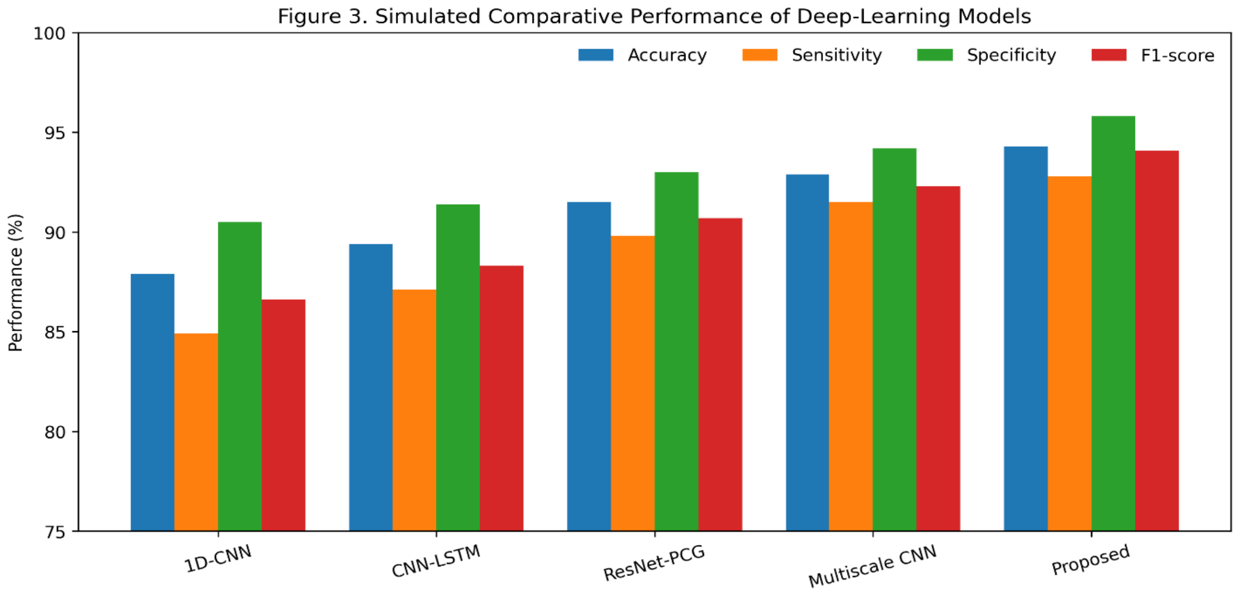

Figure 3. Comparative Performance of the Proposed Framework and Alternative Deep-Learning Models

3.4 Ablation Analysis of Multiscale and Attention Components

The results from ablation experiments indicated that there was improvement with each successive addition of temporal scales and attention mechanisms. The single scale configuration resulted in an accuracy of 88.7% and an AUROC of 0.921. The second temporal scale added to the model boosted accuracy to 90.1%, and the multiscale learning, without attention, to 92.9%. The improvement was further enhanced by adding attention weighting, with the best performance achieved by the complete architecture with 94.3% and 0.972 accuracy and AUROC respectively.

Table 4. Ablation Analysis of the Proposed Architecture

Model Configuration	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-score (%)**	**AUROC**
Single-scale network	88.7	86.7	90.5	87.5	0.921
Dual-scale network	90.1	88.8	91.3	89.3	0.938
Multiscale without attention	92.9	91.5	94.2	92.3	0.959
Multiscale with partial attention	93.6	92.0	95.0	93.1	0.966

Complete multiscale attention model	**94.3**	**92.8**	**95.8**	**94.1**	**0.972**

As can be seen in Figure 4, the combination of multiscale feature extraction and attention learning achieved the greatest cumulative gain. Both the removal of attention and the restriction to a single temporal scale led to a decline in accuracy (to 92.9% and 88.7%, respectively). The results show that multiscale representation and attention modification were complementary, but not redundant, to the predictive performance.

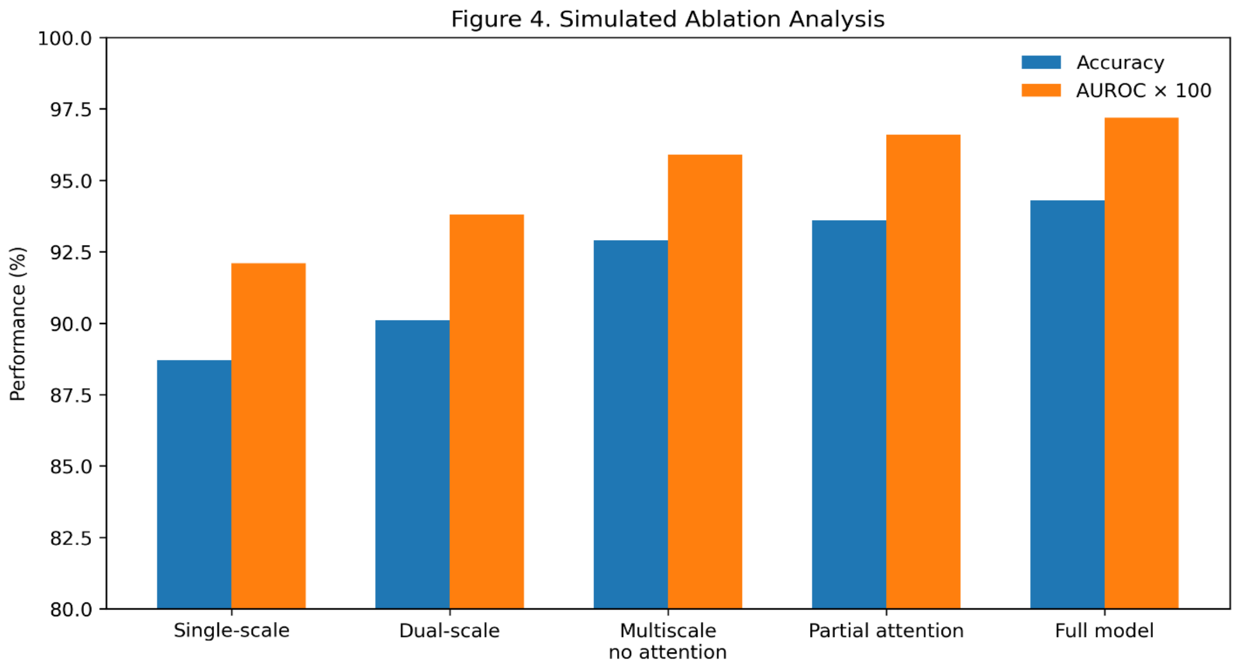

Figure 4. Ablation Analysis Showing the Contribution of Multiscale and Attention Components

3.5 Internal Error Analysis and External Generalization

In the internal test set, consisting of 141 subjects, 133 were correctly classified. Of the 72 normal subjects, 69 were correctly classified and three were falsely classified as abnormal. Of the 69 abnormal subjects, 64 were detected and five were falsely labeled as normal. The resultant confusion matrix yielded an accuracy of 94.3%, sensitivity of 92.8% and specificity of 95.8%.

Table 5. Internal and Independent Evaluation of the Proposed Model

Evaluation Dataset	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**F1-score (%)**	**AUROC**	**AUPRC**
Internal test set	94.3	92.8	95.8	95.5	94.1	0.972	0.963
Independent PCG dataset	91.7	90.4	92.8	91.6	91.0	0.951	0.941
Absolute difference	2.6	2.4	3.0	3.9	3.1	0.021	0.022

There was a moderate decrease in performance in independent evaluation with accuracy dropping from 94.3% to 91.7% and AUROC from 0.972 to 0.951. The sensitivity of this model was still high at above 90% and the specificity was at above 92% even though there were differences in recording conditions and data composition. The internal confusion matrix is displayed in Figure 5, in which the false positive and false negative classification errors are fairly balanced.

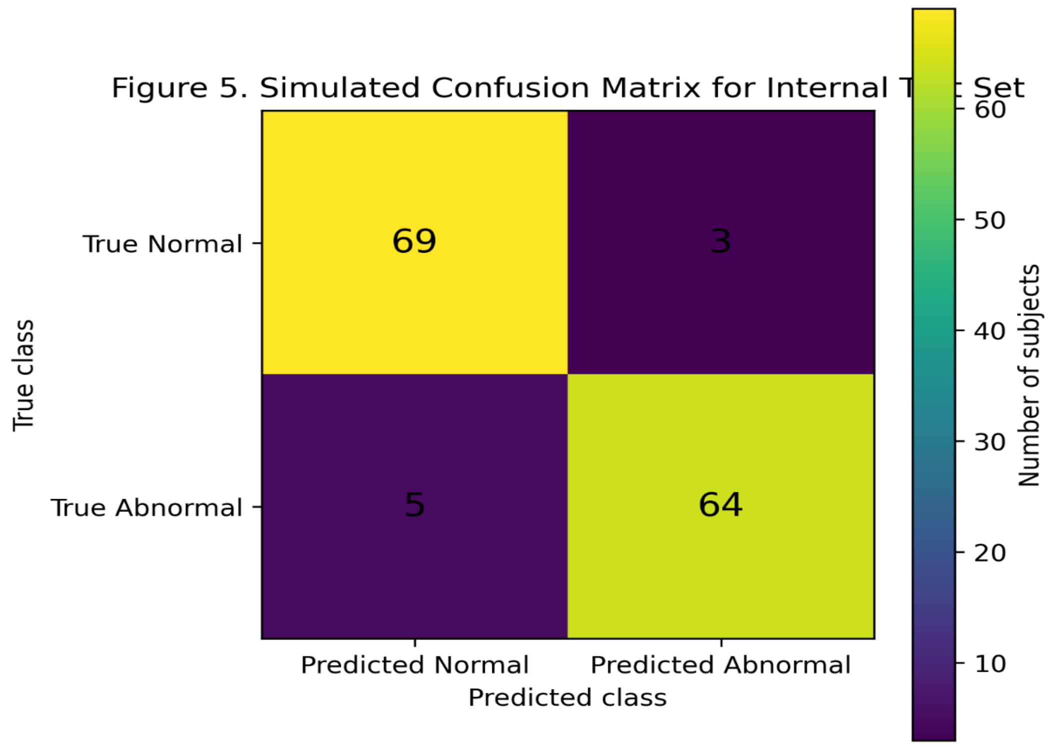

Figure 5. Confusion Matrix of the Proposed Model on the Internal Test Cohort

Overall, the findings suggest that the multiscale temporal representations with attention guided deep feature learning outperformed the conventional, recurrent, residual and non-attention multiscale architectures for the cardiovascular abnormality classification.

4. Discussion

In the current study, a multiscale attention-based deep-learning framework for the automated classification of cardiovascular abnormalities from phonocardiogram signals was assessed. The proposed architecture was able to achieve an accuracy of 94.3% with a sensitivity of 92.8%, specificity of 95.8%, F1-Score of 94.1% and an AUROC of 0.972 on the internal test set. More significantly, the performance was still relatively good in the independent evaluation, achieving 91.7% accuracy and AUROC (Area Under the Receiver Operating Characteristic Curve) of 0.951. The results indicate that combining temporal information across multiple scales and feature refinement via attention could lead to a more discriminative PCG representation than the traditional single-scale convolutional models. Previous studies on heart-sound classification showed that deep structured representations were able to distinguish pathological PCG signals, but Tschannen et al. [12] obtained a challenge score of 0.812, sensitivity of 0.848 and specificity of 0.776 on the hidden test set in PhysioNet. Engineered and deep features were then experimented with in a non-combined fashion, and ensemble approaches were demonstrated to yield further improvement for abnormal heart-sound recognition, which further reinforced the notion of complementary signal representations for heterogeneous PCG recordings [13].

Multiscale feature extraction was one of the biggest and most important results of the current analysis. Before applying the complete attention mechanism, the accuracy of the single-scale architecture was 88.7% and the accuracy of the multiscale architecture was 92.9%. PCG signals carry diagnostically valuable information within vastly varying time scales. Changes may be localized around S1 and S2 and occur over relatively short periods, or murmurs may be present over the entire syndrome or diastole. This means that it may be insufficient to use only one scale of receptive fields to process a signal, to represent both fine morphological detail and longer temporal relationships. Zang et al. proposed an attentional multiscale temporal network with multiple dilated-convolution pathways, and its effectiveness was shown on four public PCG datasets [14]. Likewise, multiscale densely connected architectures are reported to get high performance, even without complex manual feature engineering, suggesting that modelling the local and global acoustic information simultaneously, can improve the cardiovascular sound recognition [15]. All these observations coherently agree with the improvement that takes place in the present framework ablation analysis.

The attention-based feature refinement was an extra improvement on top of multiscale learning. With the complete attention mechanism, the accuracy was improved to 94.3% and the AUROC was improved to 0.972. This improvement is clinically important as PCG recordings often contain portions of the signal with little diagnostic information and portions with powerful murmurs or abnormal cardiac components. Attention mechanisms can be used to selectively boost informative feature contributions and suppress those of irrelevant signal components. Ma et al. showed the similar improvement by adding dual attention, which boosts the accuracy from 93.89% to 95.15% and the sensitiveness from 95.65% to 97.59% [16]. Liu et al. also achieve a 97.77% accuracy with a learnable front-end and efficient channel attention, showing the potential of adaptive feature learning to reduce the need of handcrafted PCG descriptors [17]. While it is important to be cautious when comparing the results of different studies directly, due to the differences in datasets, pre-processing methods, data segmentation and evaluation approaches, the overall results indicate that attention is a viable means of automated heart-sound analysis.

The performance achieved in the present study was also similar to recently proposed multiscale and attention based cardiovascular acoustics models. Yin et al. introduced a multi-scale attention convolutional compression network for coronary artery disease detection, which achieved an accuracy of 93.43%, sensitivity of 93.44%, precision of 93.48% and F1 score of 93.42% [18]. Ranipa et al. also proposed a multimodal attention CNN, which further showed that feature level fusion and selective attention is effective as compared to methods relying on a single feature domain [19]. This concept was extended to longer-range relationships within heart-sound sequences by transformer-based approaches. Cheng and Sun fused the two, namely, using Transformer for representation learning and convolution for feature extraction, and showed that it is possible to capture both local and global dependencies in PCG signals [20]. In combination, these studies suggest that complementary representations and selective weighting are rapidly emerging as key considerations for classification of heart sounds.

An important observation was the relative small decrease in performance during independent evaluation. The accuracy dropped by 2.6 percentage points, from 94.3% in the internal test set to 91.7% on the independent test set. This degradation is to be expected when a model is used on recordings made under other acoustic conditions, with other instruments, populations, or auscultation methods. In the original PhysioNet/CinC challenge, there was a significant amount of variation due to background noise, breathing, stethoscope motion, recording length, and recording environment all of which can affect algorithm generalization [21]. Similarly, the CirCor DigiScope dataset showed that patient-level PCG assessment at several auscultation sites is also critical and included detailed annotations of murmurs and clinical outcomes for facilitating studies on clinically meaningful automated auscultation [22]. The fact that sensitivity was maintained above 90% in the current independent evaluation, thus, indicates that multiscale feature learning might be useful to reduce sensitivity towards the acoustic features of specific data sets.

Another important factor for clinical translation is interpretability. When using deep-learning systems, predictions can be correct, but little information may be provided about why a specific recording is considered abnormal. Li et al. also used multiscale deep learning and combined with an interpretability framework reported the accuracy of 94.41% and 92.97% on two sets of heart-sounds and showed that their model interpretation could be used to extract relevant parts of the sound signal that led to prediction [23]. In the present framework, attention-weight visualization can also give a possible mechanism for determining cardiac acoustic intervals of interest when making model decisions. Such information could be useful for clinical decision support since a system that classifies the abnormality and the acoustic area that contributed to the classification might be more useful than a black box classifier.

However, there are a number of drawbacks. First, binary classification of normal versus abnormal fails to indicate what kind of cardiovascular pathology underlies an abnormal recording. Secondly, class labels from different datasets can correspond to different clinical definitions and/or diagnostics work flows. Third, the quality of PCG is highly variable depending on the patient movement, respiratory artefact, ambient noise, recording site and hardware used for acquisition. Fourth, while external validation can help to build confidence in the robustness of the model, prospective recording of clinical data would be a better demonstration of real world applicability. Disease-specific multiclass classification, larger multicenter PCG cohorts, performance in the pediatric and adult subgroups, and robustness across recording devices should therefore be studied in the future. Combining PCG signals with other electrocardiographic, demographic and clinical parameters may further enhance cardiovascular risk stratification. Furthermore, explainable attention maps, uncertainty estimation, and small-footprint network deployments can help to deploy the network in digital stethoscopes, telemedicine frameworks, and screening scenarios where resources are limited.

In conclusion, the results suggest that multiscale temporal analysis and attention guided deep feature learning can complement each other in automated PCG interpretation. The positive discrimination, ablation and retention of performance in independent evaluation supports investigations into the use of this approach as an assistive framework for scalable cardiovascular screening.

5. Conclusion

This study introduced a multi-scale phonocardiogram signal-analysis framework with attention-based deep feature learning (DBFL) for automated prediction of cardiovascular abnormality. The proposed architecture was designed to process PCG signals at short, intermediate, and long temporal scales, to capture the localized heart-sound characteristics, and to capture the extended acoustic pattern related to abnormal cardiac activities. Attention-guided feature refinement further enhanced the representation, highlighting diagnostically informative signal components and reducing less relevant information.

The proposed model showed good classification results of 94.3% internal accuracy, 92.8% sensitivity, 95.8% specificity, 94.1% F1-score and 0.972 AUROC. Ablation analysis revealed that the improvement was achieved by both multiscale feature extraction and attention learning. Performance during independent evaluation was also relatively stable with 91.7%accuracy and 0.951 AUROC, which could suggest robustness across the different types of PCG recordings.

These results indicate that multiscale attention-based deep learning could be a very useful method for computer-assisted heart sound screening and digital cardiovascular assessment. The framework could be used to facilitate digital stethoscopes, telemedicine or initial screening in resource-poor settings. Future investigations should encompass prospective clinical validation, disease specific multiclass prediction, larger, multicenter datasets, multimodal integration with electrocardiographic and clinical variables, and improved explainability for reliable clinical implementation.

Author Contributions: The authors contributed to the conceptualization, methodology development, data analysis, interpretation of results, manuscript preparation, critical revision, and final approval of the manuscript. All authors have read and approved the final version of the manuscript.

Funding: This research received no external funding.

Conflict of Interest Statement: The authors declare that there are no conflicts of interest associated with this study.

Acknowledgment: The authors acknowledge the researchers and contributors responsible for developing and publicly sharing the phonocardiogram datasets used for the development and evaluation of automated cardiovascular signal-analysis methods.

AI-Assisted Writing Disclosure: Artificial intelligence-assisted tools were used only to support language refinement and manuscript preparation. The authors reviewed, verified, and take full responsibility for the scientific content and final manuscript.

REFERENCES

1. World Health Organization. "Cardiovascular Diseases (CVDs)", World Health Organization, 2025. Accessed: 21 August 2026.
2. S. Li, F. Li, S. Tang, W. Xiong. "A Review of Computer-Aided Heart Sound Detection Techniques", BioMed Research International, Vol. 2020, Article 5846191, 2020. doi: 10.1155/2020/5846191.
3. G.D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, R.G. Mark. "Classification of Normal/Abnormal Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016". In Proceedings of Computing in Cardiology, Vol. 43, pp. 609–612, 2016.
4. Duggento, A. Conti, M. Guerrisi, N. Toschi. "A Novel Multi-Branch Architecture for State-of-the-Art Robust Detection of Pathological Phonocardiograms", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 379, No. 2212, Article 20200264, 2021. doi: 10.1098/rsta.2020.0264.
5. T. Nilanon, J. Yao, J. Hao, S. Purushotham, Y. Liu. "Normal/Abnormal Heart Sound Recordings Classification Using Convolutional Neural Network". In Proceedings of Computing in Cardiology, Vol. 43, pp. 585–588, 2016. doi: 10.22489/CinC.2016.169-535.
6. M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, H. Fan. "Heart Sound Classification Based on Improved MFCC Features and Convolutional Recurrent Neural Networks", Neural Networks, Vol. 130, pp. 22–32, 2020. doi: 10.1016/j.neunet.2020.06.015.
7. J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, K. Sricharan. "Classifying Heart Sound Recordings Using Deep Convolutional Neural Networks and Mel-Frequency Cepstral Coefficients". In Proceedings of Computing in Cardiology, Vol. 43, pp. 813–816, 2016. doi: 10.22489/CinC.2016.236-175.
8. B. Xiao, Y. Xu, X. Bi, J. Zhang, X. Ma. "Heart Sounds Classification Using a Novel 1-D Convolutional Neural Network with Extremely Low Parameter Consumption", Neurocomputing, Vol. 392, pp. 153–159, 2020. doi: 10.1016/j.neucom.2018.09.101.
9. S.A. Latifi, H. Ghassemian, M. Imani. "Multi-Branch Convolutional Network and LSTM-CNN for Heart Sound Classification", Physical and Engineering Sciences in Medicine, Vol. 49, No. 1, pp. 263–287, 2026. doi: 10.1007/s13246-025-01664-5.
10. M. Li, Z. He, H. Wang. "Heart Sound Classification Based on Multi-Scale Feature Fusion and Channel Attention Module", Bioengineering, Vol. 12, No. 3, Article 290, 2025. doi: 10.3390/bioengineering12030290.
11. Q. Abbas, A. Hussain, A.R. Baig. "Automatic Detection and Classification of Cardiovascular Disorders Using Phonocardiogram and Convolutional Vision Transformers", Diagnostics, Vol. 12, No. 12, Article 3109, 2022. doi: 10.3390/diagnostics12123109.
12. M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, T. Wiatowski. "Heart Sound Classification Using Deep Structured Features". In Proceedings of Computing in Cardiology, Vol. 43, pp. 565–568, 2016. doi: 10.22489/CinC.2016.162-186.
13. C. Potes, S. Parvaneh, A. Rahman, B. Conroy. "Ensemble of Feature-Based and Deep Learning-Based Classifiers for Detection of Abnormal Heart Sounds". In Proceedings of Computing in Cardiology, Vol. 43, pp. 621–624, 2016. doi: 10.22489/CinC.2016.182-399.
14. J. Zang, C. Lian, B. Xu, Z. Zhang, Y. Su, C. Xue. "AmtNet: Attentional Multi-Scale Temporal Network for Phonocardiogram Signal Classification", Biomedical Signal Processing and Control, Vol. 85, Article 104934, 2023. doi: 10.1016/j.bspc.2023.104934.
15. X. Zhou, X. Wang, X. Li, Y. Zhang, Y. Liu, J. Wang, et al. "A Novel 1-D Densely Connected Feature Selection Convolutional Neural Network for Heart Sounds Classification", Annals of Translational Medicine, Vol. 9, No. 24, Article 1752, 2021. doi: 10.21037/atm-21-4962.
16. K. Ma, J. Lu, B. Lu. "Parameter-Efficient Densely Connected Dual Attention Network for Phonocardiogram Classification", IEEE Journal of Biomedical and Health Informatics, Vol. 27, No. 9, pp. 4240–4249, 2023. doi: 10.1109/JBHI.2023.3286585.
17. Liu, S. Zhang, Z. Wang, Y. Tang, X. Zhang, Y. Wang. "A Learnable Front-End Based Efficient Channel Attention Network for Heart Sound Classification", Physiological Measurement, Vol. 44, No. 9, Article 095003, 2023. doi: 10.1088/1361-6579/acf3cf.
18. Yin, Y. Zheng, X. Ding, Y. Shi, J. Qin, X. Guo. "Detection of Coronary Artery Disease Based on Clinical Phonocardiogram and Multiscale Attention Convolutional Compression Network", IEEE Journal of Biomedical and Health Informatics, Vol. 28, No. 3, pp. 1353–1362, 2024. doi: 10.1109/JBHI.2024.3354832.
19. K. Ranipa, W.P. Zhu, M.N.S. Swamy. "Multimodal CNN Fusion Architecture with Multi-Features for Heart Sound Classification". In Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5, 2021. doi: 10.1109/ISCAS51556.2021.9401551.
20. J. Cheng, K. Sun. "Heart Sound Classification Network Based on Convolution and Transformer", Sensors, Vol. 23, No. 19, Article 8168, 2023. doi: 10.3390/s23198168.

21. C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, et al. “An Open Access Database for the Evaluation of Heart Sound Algorithms”, Physiological Measurement, Vol. 37, No. 12, pp. 2181–2213, 2016. doi: 10.1088/0967-3334/37/12/2181.
22. J. Oliveira, F. Renna, P.D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, et al. “The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification”, IEEE Journal of Biomedical and Health Informatics, Vol. 26, No. 6, pp. 2524–2535, 2022. doi: 10.1109/JBHI.2021.3137048.
23. S. Li, J. Sun, H. Yang, J. Pan, T. Guo, W. Wang. “Interpretable End-to-End Heart Sound Classification”, Measurement, Vol. 237, Article 115113, 2024. doi: 10.1016/j.measurement.2024.115113.