# Multistage Fusion Framework for Coronary Artery Disease Detection from Multichannel PCG

Arnab Maity[1], Souvik Sinha[1], Matthew Fynn[2], Milan Marocchi[2], Kayapanda Mandana[3], Yue Rong[2], and Goutam Saha[1]

[1] Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India
arnab10maity@iitkgp.ac.in, souvik_sinha@kgpian.iitkgp.ac.in, gsaha@ece.iitkgp.ac.in

[2] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, Australia
matthew.fynn@postgrad.curtin.edu.au, milan.marocchi@postgrad.curtin.edu.au, y.rong@curtin.edu.au

[3] Department of Cardiology, Fortis Healthcare, Kolkata, 7007107, West Bengal, India
kmmandana@gmail.com

**Abstract.** Coronary artery disease (CAD) remains one of the leading causes of mortality worldwide. Phonocardiogram (PCG) signals offer a non-invasive, affordable, and accessible means for early detection of CAD. However, the diverse acoustic manifestations of the disease across different auscultation sites make accurate diagnosis using a single-channel stethoscope challenging. Moreover, the scarcity of large annotated datasets further limits the development of robust diagnostic models. This work presents a multichannel CAD detection framework using transfer learning that leverages both early/late fusion from multiple auscultation sites. A lightweight pretrained deep learning model is designed to address data scarcity and enable computationally efficient deployment. We explore early and late fusion strategies to extract the channel-wise collective information in detecting the CAD. The proposed system achieves a 9.46% improvement in accuracy over its single-channel counterpart, highlighting its potential for practical and scalable CAD screening. Clinically, it provides an affordable, accessible, and efficient tool for CAD detection, especially in low-resource settings.

**Keywords:** Coronary artery disease · Phonocardiogram · Transfer learning · Embedding fusion · Multichannel Stethoscope

## 1 Introduction

Cardiovascular diseases (CVDs) represent the foremost global health concern, accounting for approximately 31% of deaths worldwide [14]. Among the diverse spectrum of CVDs, coronary artery disease (CAD) is the primary contributor to CVD-related deaths and often serves as a precursor to other life-threatening diseases. Early detection of CAD is crucial to mitigate its progression and associated complications. Conventional diagnostic techniques such as coronary angiography require specialized medical infrastructure, are expensive, and invasive [1]. Angiography is typically reserved for symptomatic patients. However, CAD can be present without symptoms, highlighting the need for a global pre-screening tool for early detection. Heart auscultation is a cost-effective preliminary screening tool for detecting CAD by identifying abnormal heart murmurs, which are due

to turbulent blood flow caused by coronary artery blockages. Phonocardiogram (PCG) graphically represent heart sounds and reflect cardiac activity, which can change with disease [1].

Early research in CAD detection primarily focused on handcrafted and nonlinear features [6, 11]. These spectral-based studies often assumed heart sound signals to be stationary, which contradicts the literature. With advancements in deep learning, researchers have shifted towards automated feature extractors using convolutional neural networks (CNN) [9]. The need for extensive feature engineering was eliminated by extracting representations directly from raw signals [6, 4, 10]. Early studies on CVD detection predominantly relied on single-channel PCG signals, using only one auscultation site to predict disease signatures. However, recent research indicates that analyzing data from multiple auscultation sites is crucial for identifying robust disease markers [8, 13, 5, 4]. Furthermore, studies utilizing multi-channel PCG have explored advanced features, including entropy-based features and cross-entropy analysis [5, 10]. These investigations collectively underscore the importance of dataset diversity in multichannel PCG. Pathak et al. [10] applied the synchrosqueezing transform to extract entropy features using four stethoscopes and used a support vector machine classifier to detect the CAD. Zhao et al. [15] proposed a hybrid convolution transformer neural network for extracting local features. However, the method relies on a separate algorithm to segment the PCG signals, adding preprocessing overhead and deployment constraints. Fynn et al. [1] proposed a seven-channel, wearable, non-invasive vest-based data acquisition system that enables enhanced data collection without the need for special assistance, used linear frequency cepstral coefficient (LFCC) features, and achieved an accuracy of 80.44%, however, relied on manual PCG signal segmentation. Therefore, current CAD detection methods often depend on cardiac cycle segmentation, requiring extensive preprocessing and occasionally human oversight. Deep learning-based methods often involve computationally intensive transformations that increase complexity. While multichannel techniques have a strong potential, current literature typically selects optimal channel combinations based on test results, limiting real-world applicability.

This study presents a novel approach for CAD detection using transfer learning to extract embeddings from multichannel PCG signals recorded at seven distinct auscultation sites. Instead of relying on computationally intensive cardiac cycle segmentation, we extract fixed-length fragments directly from time-frequency (TF) representations. A lightweight, pretrained deep learning model is used for embedding extraction, reducing system complexity while maintaining performance. To ensure unbiased and generalizable results, embedding-level fusion is performed by combining channels based on their individual validation performance. Furthermore, score-level fusion of predictions from various channel combinations is applied to reinforce the effectiveness of multichannel integration.

## 2 Database description

Seven-channel PCG data was collected using a wearable vest equipped with multiple stethoscopes, worn by each subject to record PCG signals from vari-

ous auscultation sites [1]. A pictorial representation of the stethoscopes used for data collection is shown in Figure 1. The data was acquired at Fortis Hospital, Kolkata, India. For this study, 60 s recordings were obtained in noisy hospital environment from 71 normal and 119 CAD male patients in a seated position who were breathing normally, excluding individuals with valvular pathologies. Diagnosis was confirmed using coronary angiography. The mean (standard deviation) age of CAD and normal patients was 60(15) years and 50(10) years.

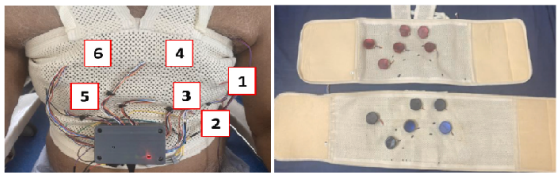


**Fig. 1.** Wearable vest during data collection (left) and vest fitted with electronic stethoscopes (right) [1].

## 3 Methodology

### 3.1 Preprocessing and feature extraction

We applied an 8th-order Butterworth bandpass filter with cutoff frequencies of 25 Hz and 400 Hz to remove low and high-frequency noise, respectively [7]. This was followed by z-score normalization across all channels to standardize amplitude variations and mitigate inter-recording variability. To extract discriminative TF features that capture the non-stationary nature of PCG signals, we employed a log mel spectrogram and a continuous wavelet transform (CWT) based scalogram (Figure 2). For the mel-spectrogram, we use 64 mel filter banks, and for the scalogram, we employ the Morlet (Gabor) wavelet as the mother wavelet [10].

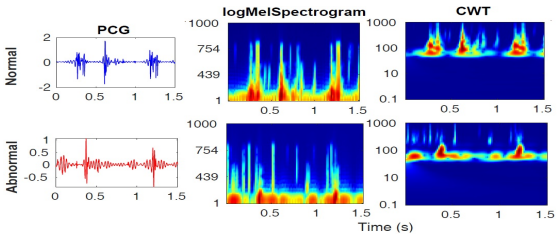


**Fig. 2.** Time-frequency representations of PCG signals for normal and CAD cases.

### 3.2 Classification framework

The YAMNet model [12] is an efficient, lightweight deep learning architecture specifically designed for audio classification. It employs depthwise separable convolution layers based on the MobileNet architecture and significantly reduces computational complexity and model size (15.5 MB) [12, 3]. It is pre-trained on a vast corpus of 1.2 million YouTube audio segments (AudioSet), provides a rich repository of acoustically relevant features [2, 6]. In this work, transfer learning with the YAMNet architecture is used to overcome the inherent challenges of low-resource datasets, primarily the risk of overfitting and the inability to learn robust, generalizable features.

**Early fusion:** Since multichannel heart-sound recordings are collected from different auscultation sites, they capture diverse information reflecting the functional characteristics of the human heart. We hypothesize that embeddings extracted from these channels contain discriminative information that can be leveraged for CAD detection. Thereby, we propose an early embedding fusion approach with deep feature representations extracted from each channel. The input $x \in \mathbb{R}^{C\times T\times F}$ consists of TF representations from $C$ channels, where $T$ and $F$ represent the time frames and frequency bins, respectively. For each selected channel $f_i = x_i \in \mathbb{R}^{1\times T\times F}$, features are extracted using a frozen YAMNet backbone up to its penultimate convolutional block. The convolution block consists of the last five convolutional layers of the YAMNet architecture, specifically those following the frozen layers and preceding the final fully connected (FC) layer. Each channel's output is processed through a global average pooling (GAP) layer followed by flattening, yielding embeddings $e_i \in \mathbb{R}^D$, where $D = 1024$. These embeddings are then fused by computing their mean across all $n$ selected channels, resulting in a single representation $E = \frac{1}{n}\sum_{i=1}^{n} e_i, \;\; E \in \mathbb{R}^D$. Finally, $E$ is passed through a common FC layer to produce the output prediction $\hat{y} \in \mathbb{R}^K$, where $K$ is the number of target classes. The convolutional block, along with the FC layer, is unfrozen for fine-tuning. Selective unfreezing helps adapt high-level features for CAD detection while retaining YAMNet's general acoustic knowledge [6]. The mean is used instead of concatenation to control model complexity and avoid increasing feature dimensionality. This strategy preserves the original embedding size, keeps the classification lightweight, and reduces the number of trainable parameters. Also, averaging promotes more robust representations by reducing overfitting due to limited data. For single-channel operation, there will be no fusion. A schematic representation of this method is shown in Figure 3.

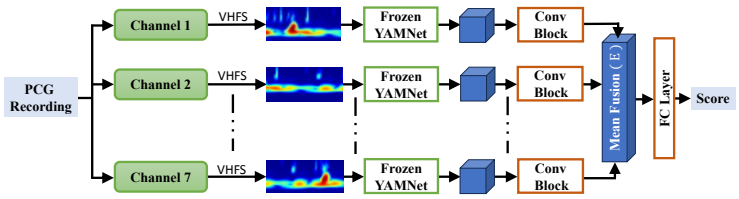


**Fig. 3.** Block diagram of the multi-channel embedding fusion using the YAMNet model.

**Late fusion:** We further enhance diagnostic robustness through a late score-level fusion strategy. The goal is to combine outputs from multiple model configurations trained on different subsets of input channels ($S_c \subseteq \{1, 2, \ldots, 7\}, \; S_c \neq \emptyset$) at the decision level to avoid the risk of misclassification by any single configuration. Let $\mathcal{S} = [S_1, S_2, \ldots, S_n]$ denote the set of selected channel subsets, and let $s_{[S_c]} \in \mathbb{R}^K$ represent the confidence score vector produced by the model trained on subset $S_c$, where $K$ is the number of classes. The final fused prediction is computed as $s_{\text{fused}} = \frac{1}{n}\sum_{c=1}^{n} s_{[S_c]} \in \mathbb{R}^K$. This approach aggregates complementary information and provides a more reliable and generalized prediction.

**Input preparation:** CNNs typically require fixed-size inputs for effective feature extraction. However, TF representations of PCG signals vary in length due

**Table 1.** Channel-wise performance comparison across different features

| Channel | log-mel spectrogram | | | Scalogram | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | Sens. | Spec. | Acc. | Sens. | Spec. | F1 score | UAR | Valid. Acc. | Rank |
| 1 | 59.96 | 78.15 | 28.96 | 64.17 | 79.83 | 37.7 | 73.46 | 58.81 | 65.74 | 4 |
| 2 | 59.96 | 74.79 | 35.06 | 58.37 | 70.59 | 38.05 | 67.01 | 54.32 | 67.35 | 3 |
| 3 | 61.56 | 73.11 | 41.82 | **67.86** | 81.51 | **44.94** | 76.11 | **63.22** | **70.99** | 1 |
| 4 | **71.03** | **85.71** | 46.36 | 66.84 | **88.24** | 30.78 | **76.68** | 59.51 | 68.93 | 2 |
| 5 | 60.54 | 73.95 | 38.31 | 61.58 | 74.79 | 39.48 | 70.45 | 57.14 | 63.13 | 6 |
| 6 | 63.68 | 73.95 | **46.75** | 64.76 | 81.51 | 36.75 | 74.38 | 59.13 | 62.07 | 7 |
| 7 | 61.55 | 73.11 | 42.34 | 62.09 | 79.83 | 32.08 | 71.34 | 55.95 | 65.25 | 5 |

to differences in signal duration. Prior studies addressed this using cardiac cycle-wise segmentation with zero-padding. Extracting individual cycles demands a separate algorithm, adding complexity to preprocessing. This work segments the TF matrix into fragments of uniform length using Variable Hop Fragment Selection (VHFS) proposed in [7]. VHFS extracts multiple overlapping fragments of fixed duration (2.5 s), covering at least one full cardiac cycle, from the TF representation of PCG signals by varying the hop length. The number of fragments per sample is determined by the fragment selection factor ($n_f$), ensuring balanced class distributions and increased training data. VHFS bypasses the computationally intensive and error-prone step of precise heart sound localization and segmentation, reducing system complexity and computational overhead. For both validation and test datasets, an equal number of fragments is selected from each sample, regardless of its length or label.

## 4 Results and discussion

The accuracy, sensitivity, specificity, F1 score, and unweighted average recall (UAR) are calculated to assess the system's performance [7]. We implemented the stratified 7-fold cross-validation by splitting the dataset into seven distinct folds. We perform the experiment with Adam optimizer over (30 epochs, batch size 64, and learning rate 0.001). To avoid overfitting, we implement an early stopping criterion with a patience of five epochs. To evaluate subject-level performance, a majority voting is applied to the predictions of all fragments belonging to a subject. The label assigned to the subject is determined by the most frequently predicted label among its fragments.

### 4.1 TF feature-based performance comparison on single channel

Table 1 presents the channel-wise performance comparison for two input TF features. The scalogram achieves 3.36% higher average sensitivity across channels than log-mel spectrograms. Scalograms provide a richer TF representation by preserving both fine-grained temporal and spectral variations, which are critical in detecting subtle pathological cues in PCGs. In all further experiments, we have considered the scalogram as an input feature to the model. To analyze channel relevance, we ranked the channels based on their validation performance, ensuring the most informative channels are emphasized. This approach guides the multi-channel feature fusion. Validation accuracy was chosen instead of test accuracy to ensure unbiased evaluation targeting real-time implementation. Channel 3 stands out as the most informative, with the highest validation

**Table 2.** Performance comparison for embedding level fusion of multiple channels based on different combination types

| Combn. type | Ch. combn. | Validation | | | Test | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc. | Sens. | Spec | Acc. | Sens. | Spec. | F1 score | UAR |
| Type I | **[3,4]** | **75.23** | 87.39 | 54.68 | **75.23** | **88.24** | 52.99 | 81.79 | **70.61** |
| | [3, 4, 2] | 73.09 | 84.87 | 53.25 | 66.23 | 75.63 | 50.26 | 73.52 | 62.94 |
| | [3, 4, 2, 1] | 72.54 | 84.03 | 53.12 | 69.46 | 80.67 | 50.91 | 76.72 | 65.79 |
| | [3, 4, 2, 1, 7] | 71.54 | 82.35 | 53.12 | 71.52 | 83.19 | 51.82 | 78.10 | 67.51 |
| | [3, 4, 2, 1, 7, 5] | 74.70 | 86.55 | 54.55 | 69.01 | 84.87 | 42.47 | 77.26 | 63.67 |
| | [3, 4, 2, 1, 7, 5, 6] | 66.86 | 77.31 | 49.48 | 74.42 | 63.40 | 71.01 | **83.19** | 50.26 |
| Type II | [2, 4] | 70.46 | 82.35 | 50.52 | 67.35 | 79.83 | 46.36 | 75.24 | 63.10 |
| | [2, 4, 3] | - | - | - | - | - | - | - | - |
| | [2, 4, 3, 6] | 66.25 | 80.67 | 42.08 | 68.41 | 78.15 | 51.95 | 75.47 | 65.05 |
| Type III | **[3,4]** | - | - | - | - | - | - | - | - |
| Type IV | [1, 3, 7] | 74.17 | **88.24** | 50.39 | 67.88 | 81.51 | 45.06 | 76.18 | 63.29 |
| Type V | [3, 4, 5, 6, 7] | 63.61 | 67.23 | **57.40** | 68.94 | 62.31 | **67.84** | 71.43 | 61.69 |

accuracy (70.99%), test accuracy (67.86%), and UAR (63.22%), indicating its superior ability to capture diagnostically relevant patterns.

### 4.2 Embedding level fusion of multiple channels

We now perform embedding-level fusion across multiple channels, based on individual channels' ranking indicated in Table 1. Different auscultation sites across the chest exhibit site-specific acoustic characteristics, providing a multi-view perspective on pathological patterns. Table 2 represents the performance of the different channel combinations. Multi-channel fusion consistently outperforms single-channel models, highlighting the value of integrating information from different auscultation sites. The combination of channels 3,4 has the highest impact with 75.23% test accuracy. However, not all combinations lead to improved performance, as seen with $[3, 4, 2]$ and $[2, 4]$. This highlights the complexity of selecting the optimal set of PCG channels. With seven available channels, there are $(2^7 - 1) - 7 = 120$ non-trivial combinations (excluding single channels), making exhaustive evaluation impractical in real-world scenarios. We categorized the channel combinations into different types to guide the selection process. We selected a representative subset of combinations for further analysis based on the following criteria: **Type I**: Based on single channel ranking (Table 1), **Type II**: Based on the anatomical and physiological relevance of channels, **Type III**: Based on highest validation accuracy, **Type IV**: Based on highest validation sensitivity, and **Type V**: Based on highest validation specificity. Type I channel combinations are the most effective, achieving higher performance compared to single channels. Type II combinations, selected based on the clinical relevance of auscultation sites, partially overlap with Type I, further reinforcing the clinical significance of the selected channels. Type III, Type IV, and Type V combinations are included for score-level fusion strategies in subsequent stages.

### 4.3 Score level fusion of single/multiple channels combinations

We integrate the output scores of individual models trained on different channel combinations to further enhance the insight. We used a selected channel combination from the combination types to perform the score-level fusion as indicated in Table 2. Table 3 demonstrates that the score-level fusion strategy

achieved a 2.09% better accuracy and 2.93% better UAR over feature-level fusion alone, particularly when using top channel combinations from all five types [3, [34], [24], [137], [34567]]. The sequential application of early (embedding-level) and late (score-level) fusion enhances the interpretability of the model and reinforces the diagnostic potential of multichannel PCG analysis.

**Table 3.** Score level fusion on the different single and multi-channel combinations

| Criterion | Ch. combns. | Acc. | Sens. | Spec. | F1 | UAR |
|---|---|---|---|---|---|---|
| Single ch. | [3, 4] | 70.48 | **90.76** | 36.36 | 79.43 | 63.56 |
| | [2, 3, 4] | 68.90 | 88.24 | 36.49 | 77.78 | 62.37 |
| Multi ch. | [[34], [24], [137], [34567]] | 73.64 | 84.03 | 55.97 | 79.82 | 70.01 |
| | [[34], [24], [234], [137], [34567]] | 75.21 | 85.72 | 57.40 | 81.25 | 71.56 |
| Single & multi ch. | [3, [34], [24], [137], [34567]] | **77.32** | 88.24 | **58.83** | **82.93** | **73.54** |

Signals from the left fourth and second intercostal (IC) spaces (channels 3 and 4) are most informative. Fusion across channels enhanced disease classification, with feature-level fusion capturing local patterns and score-level (late) fusion further boosting performance. Also, the use of a computationally lightweight model supports the feasibility of deployment in portable or edge devices. However, the relatively low specificity observed indicates a tendency toward false positives, likely due to limited variability in normal signals. The data was collected in a real-world hospital environment, where substantial ambient noise and operational constraints affected signal quality. In particular, the wearable sensor vest was not optimally fitted for each patient, leading to inconsistent contact and motion-induced artifacts. This suggests a need for better data balancing and adaptive denoising in future studies.

## 5 Conclusion

This study proposes a robust framework for classifying CAD using multichannel PCG signals. Two TF feature representations, log-mel spectrogram and scalogram, are used to extract discriminative patterns from the signals. Pretrained YAMNet model in a transfer learning setup served as a lightweight embedding extractor. We first performed embedding-level fusion by combining representations from multiple PCG channels. Then, we applied score-level fusion by integrating prediction scores from models with single and multi-channel inputs. The proposed system achieved a 7.37% gain in accuracy over the best single-channel model by fusing multichannel embeddings, underscoring the benefit of using heart sounds from multiple auscultation sites. Building on this, the addition of score-level fusion further enhanced the accuracy by an additional 2.09%. The proposed system achieved an accuracy of 77.32%, sensitivity of 88.24%, and UAR of 73.54%. In future work, controlled data collection, more advanced feature extraction, and model architectures can be explored to improve generalizability further.

**Acknowledgement:** This work is funded by The Scheme for Promotion of Academic and Research Collaboration (SPARC), Indian Institute of Technology Kharagpur, and is done in collaboration with Curtin University, Perth, Australia. We thank Ticking Heart Pty Ltd for use of their wearable vest for data collection.

## References


1. Fynn, M., Mandana, K., Rashid, J., Nordholm, S., Rong, Y., Saha, G.: Practicality meets precision: Wearable vest with integrated multi-channel PCG sensors for effective coronary artery disease pre-screening. Computers in Biology and Medicine **189**, 109904 (2025)
2. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. pp. 776–780. IEEE (2017)
3. Hershey, S., Chaudhuri, S., Ellis, D.P., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 131–135. IEEE (2017)
4. Huang, Q., Yang, H., Zeng, E., Chen, Y.: A deep-learning-based multi-modal ECG and PCG processing framework for label efficient heart sound segmentation. In: 2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 109–119. IEEE (2024)
5. Li, H., Wang, X., Liu, C., Zeng, Q., Zheng, Y., Chu, X., Yao, L., Wang, J., Jiao, Y., Karmakar, C.: A fusion framework based on multi-domain features and deep learning features of phonocardiogram for coronary artery disease detection. Computers in Biology and Medicine **120**, 103733 (2020)
6. Maity, A., Pathak, A., Saha, G.: Transfer learning based heart valve disease classification from phonocardiogram signal. Biomedical Signal Processing and Control **85**, 104805 (2023)
7. Maity, A., Saha, G.: Time-frequency fragment selection for disease detection from imbalanced phonocardiogram data. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society. pp. 1–4. IEEE (2023)
8. Makaryus, A.N., Makaryus, J.N., Figgatt, A., others.: Utility of an advanced digital electronic stethoscope in the diagnosis of coronary artery disease compared with coronary computed tomographic angiography. The American Journal of Cardiology **111**(6), 786–792 (2013)
9. Megalmani, D.R., Shailesh, B., Rao, A., Jeevannavar, S.S., Ghosh, P.K.: Unsegmented heart sound classification using hybrid CNN-LSTM neural networks. In: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 713–717. IEEE (2021)
10. Pathak, A., Mandana, K., Saha, G.: Ensembled transfer learning and multiple kernel learning for phonocardiogram based atherosclerotic coronary artery disease detection. IEEE Journal of Biomedical and Health Informatics (2022)
11. Pathak, A., Samanta, P., Mandana, K., Saha, G.: Detection of coronary artery atherosclerotic disease using novel features from synchrosqueezing transform of phonocardiogram. Biomedical Signal Processing and Control **62**, 102055 (2020)
12. Plakal, M., Ellis, D.: Yamnet. [Online: Accessed 15 June 2025] Available: https://github.com/tensorflow/models/tree/master/research/audioset/ (2021)
13. Schmidt, S.E., Holst-Hansen, C., Hansen, J., Toft, E., Struijk, J.J.: Acoustic features for the identification of coronary artery disease. IEEE Transactions on Biomedical Engineering **62**(11), 2611–2619 (2015)
14. World Health Organization: Cardiovascular diseases (CVDs). [online: Accessed 15 June 2025]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2021)
15. Zhao, W., Ma, H., et al.: Detection of coronary heart disease based on heart sound and hybrid vision transformer. Applied Acoustics **230**, 110420 (2025)