Bimodal ECG-PCG Cardiovascular Disease Detection: a Close Look at Transfer Learning and Data Collection Issues

Alessia Calzoni[1,2,*], Mattia Savardi[3] and Alberto Signoroni[3,*]

[1]Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy
[2]Isinnova SRL, Via Berlinguer, 2, 25124, Brescia, Italy
[3]Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, Viale Europa 11, 25121, Brescia, Italy

Abstract

Early detection of cardiovascular diseases (CVDs) is crucial for minimizing their adverse impact on patients' health. Electrocardiograms (ECGs), which capture the heart's electrical activity, have been widely used to primarily evaluate heart conduction disorders. On the other hand, phonocardiograms (PCGs) recorded during cardiac auscultation, have been less explored, often being overlooked in favor of echocardiograms for detecting mechanical issues such as valvular diseases. However, due to their low cost and non-invasive nature, the analysis of both ECGs and PCGs can be easily integrated into preventive settings. Combining effectively the complementary information from these two modalities could significantly enhance the early detection of CVDs, where Machine Learning (ML) techniques can offer promising and cost-effective solutions. Progress in this area, however, has been limited by the lack of large enough datasets containing both ECG and PCG signals. One objective of this work is to analyze in-depth prior bimodal CVD detection research, identifying key issues to better address data collection and transfer learning limitations. We also propose a different approach to transfer learning for improving heart sound interpretation. Our findings confirm the effectiveness of using both signals to detect abnormal heart conditions. However, we also notice that even a refined transfer learning approach to enhance PCG interpretation is not enough to fully address the issues coming from the lack of bimodal data, indicating the need for further efforts in this direction. Ultimately, our bimodal approach achieved an overall AUROC of 96.4%, exceeding the performance of corresponding ECG-only and PCG-only models by approximately 3% and 10%, respectively. Compared to the other existing approaches, our method demonstrated superior AUROC performance while maintaining a relatively low false-negative rate, which is critical in CVD screening contexts.

Keywords

Cardiovascular diseases, electrocardiogram, phonocardiogram, transfer learning, multi-modality

1. Introduction

Cardiovascular diseases (CVDs) are the primary cause of morbidity and mortality worldwide, accounting for an estimated 17.9 million deaths each year, decreasing the quality of life and imposing a subsequent significant burden on global healthcare systems [1]. Early diagnosis systems play a crucial role in mitigating the negative effects of these diseases, as delayed identification is a major reason for high morbidity and death [2].

Given the complexity of heart activity, CVDs encompass different conditions, such as arrhythmias, valvular diseases, and coronary artery disease. Clinicians use a range of diagnostic tools, including echocardiography, computer tomography scans, and angiography: all of these approaches, although highly specific and considered gold standards for the diagnoses, turn out to be expensive and not accessible in primary care, limiting their use for large-scale screening [3, 4]. In contrast, electrocardiography and cardiac auscultation are widely used methodologies for early detection of pathological cardiac conditions due to their low cost, non-invasiveness, and ease of measurement, guiding for further specific diagnostic examinations [2]. An electrocardiogram (ECG) records the heart's electrical activity during each heartbeat, while a phonocardiogram (PCG), the audio signal obtained via an electronic stethoscope during heart auscultation, captures the vibrations caused by the flow of blood through the

3rd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2024), November 25–28, 2024, Bolzano, Italy
*Corresponding author.
alessia.calzoni@unibs.it (A. Calzoni); alberto.signoroni@unibs.it (A. Signoroni)

heart's chambers. These vibrations are generated by the mechanical activity of the heart valves as they open and close, which occurs in response to the electrical activation that triggers atrial and ventricular contractions during the cardiac cycle [5]. As shown in Figure 1, due to their physiological origin, the two signals, despite very different, are strictly related. Although ECGs and PCGs signals are widely used to detect CVDs, studies have shown that interpretation accuracy can vary significantly depending on the clinician's level of training, with expert cardiologists achieving a median accuracy of less than 80% for both signals [6, 7]. Currently, neither the ECG nor the PCG signal alone can fully diagnose CVDs, as these two modalities reflect the heart's electrical and mechanical activities, respectively [3]. Heart sounds can reveal pathological conditions of the cardiac valves, while ECG is better suited for detecting conduction disorders: in particular, different patients may exhibit similar heart sounds but different ECGs, or vice versa [8].

In this scenario, Machine Learning (ML) methods can be leveraged to develop automatic diagnostic tools that could enhance early detection of CVDs in a preventive setting. So far, most efforts have focused mainly on single-modality approaches, using either ECG [9, 10, 11] or PCG [12, 13, 14] signals [3]. Most approaches that use ECG signals have been developed to detect conduction or structural issues, as publicly available datasets typically contain these labels. Due to the electrical nature of the ECG signal, conditions that are primarily or more effectively identified through cardiac auscultation (such as valvular diseases) have been less studied and modeled using ECG signals. Moreover, this context lacks public datasets, and the few studies that utilize ECGs to identify valvular diseases are based on non-publicly available datasets [4], making the PCG counterpart preferred to investigate these conditions. Recent research explores multimodal techniques to improve diagnostic performance, taking advantage of complementary information given by the two signals [3]. The limited exploration in this area is primarily due to the lack of a large dataset with combined PCG and ECG recordings. There is only one publicly available bimodal dataset from both healthy and pathological individuals, that enables the development of a classification model for abnormal heart conditions. Actually, there is another open-access database of synchronous ECGs and PCGs that consists only of signals from healthy subjects undergoing different stress-test sessions, making it unsuitable for pathological classification tasks [15]. Due to strict regulations, data collection in the healthcare field is significantly more challenging than in other fields, such as computer vision. This lack of data can affect the development of accurate ML models, especially in the supervised paradigm, where data labeling is often costly and time-consuming. In this context, transfer learning has arisen as a valuable approach, often demonstrating the ability to improve performance, mitigating overfitting issues commonly faced by models built from scratch on limited datasets. With transfer learning, models pre-trained on large datasets can be leveraged to "transfer knowledge" across different domains. These pre-trained models can serve as feature extractors or be further fine-tuned on the data of the new task. The core principle is to transfer knowledge acquired in one domain (source) and apply it to a new task in a related but different domain (target) [16, 17]. Although transfer learning has already been applied to address data limitations in detecting abnormal heart conditions using bimodal approaches, its full potential remains underexplored, and several possible improvements have not been investigated. Based on existing findings, interpreting PCG signals has proven to be more challenging compared to the electrical counterpart. However, not all publicly available PCG unimodal datasets have been fully leveraged to enhance this signal comprehension. Additionally, since PCG signals are essentially audio recordings, employing audio data as the source domain in a transfer learning setting could potentially improve PCG interpretation and, by extension, enhance overall bimodal performance. Therefore, in this work, we aim to investigate whether applying transfer learning from audio recordings, along with the use of a larger amount of unimodal PCG data, could effectively enhance the understanding of heart sounds, and improve the ability to detect CVDs by combining the information coming from ECGs.

The main contributions of this work can be summarized as follows:

- analyzing in-depth the literature, deriving observations and highlighting problems that are not always evident in individual works, trying to explore some of their limitations better;
- investigating the potential of a transfer learning approach from audio data to improve abnormality

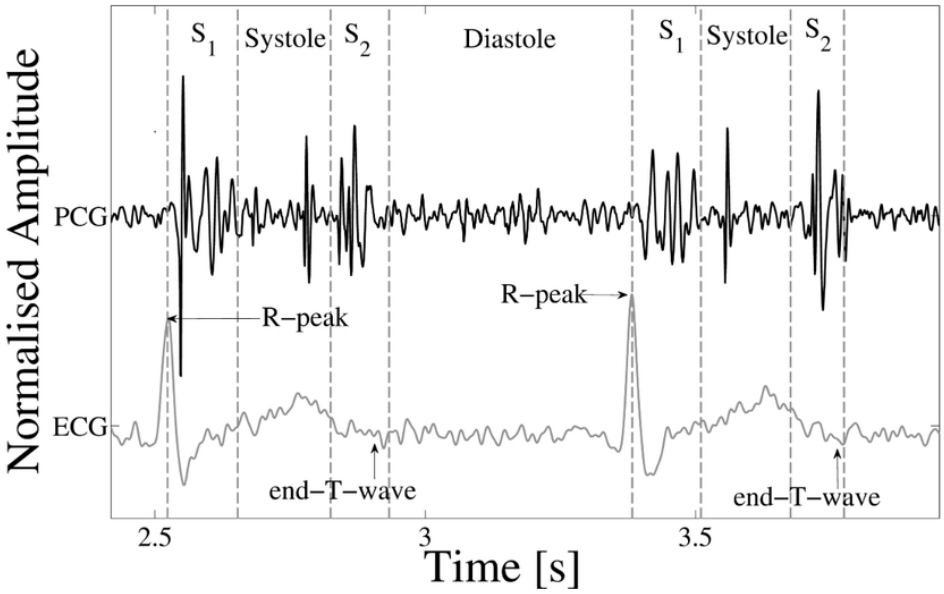

Figure 1: ECG-PCG relationship: the first (S_1) and second (S_2) heart sounds occur at the beginning of the systole and diastole phases, respectively, as a result of the closure of the mitral and tricuspid valves in the first case, and the aortic and pulmonary valves in the latter [5].

detection in PCG recordings, leveraging all the unimodal PCG datasets for the fine-tuning process;
- developing a bimodal model based on both ECG and PCG signals and confirming the effectiveness of multimodal analysis compared to single-modality approaches.

2. Related Work

The first attempt to fully leverage the information underlying ECG and PCG signals to improve the identification of chronic and non-conduction heart disorders dates back to 2019 [18]. Using simultaneously collected ECG and PCG recordings, a novel dual-input neural network was developed, integrating both conventional feature extraction and deep learning. That work proved for the first time that combining both signals significantly improves performance in detecting coronary artery disease compared to analyzing just one. Traditional methods, such as support vector machine, were employed by Chakir et al., Singh et al., and Li et al., using manually extracted features from synchronized ECG and PCG signals to classify normal and abnormal heart conditions [19, 20, 21]. Within this context, other studies have combined different decomposition techniques as feature extractors, which are then used as inputs for neural networks [22, 23]. To overcome the limitations of manual feature extraction required in traditional machine learning methods, several studies have leveraged deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) networks. By automatically extracting hierarchical features directly from raw data and optimizing complex representations, deep learning models eliminate the need for manual feature engineering, and different configurations have been investigated in the classification of CVDs [3, 24, 25, 26]. To tackle the challenge of limited bimodal data, the first transfer learning approach was introduced by Hettiatachchi et al. in 2021 [27]. Time-frequency representations of both ECG and PCG signals were used as inputs of two CNNs initially trained separately on publicly available unimodal datasets of ECGs and PCGs. The results showed that the performance on bimodal data significantly improved when using the hybrid model using these pre-trained networks to extract more meaningful and representative features, compared to PCG-only or ECG-only models. A second transfer learning approach utilized two slightly modified VGG-16 architectures [28] for the two branches of ECG and PCG signals, with a 2-dimensional time-frequency transformation as input. The study explored two strategies: using pre-trained ImageNet models as

feature extractors and fine-tuning them using an ECG and a PCG unimodal dataset for the two branches, obtaining better results in the first case [29]. These latter transfer learning approaches rely on a single publicly available PCG dataset, limiting the potential improvements that could come from incorporating additional data sources. However, with other unimodal PCG datasets now available, there is an opportunity to enhance heart sound interpretability, which often appears to be the limiting factor in bimodal models.

3. Datasets

In this section, we provide a detailed description of the most significant and widely used unimodal PCG datasets, focusing on key aspects such as data sources, recording conditions, and label characteristics. Additionally, we thoroughly examine the bimodal dataset, which contains simultaneous recordings of both PCG and ECG signals from healthy and pathological individuals, used in the development of the bimodal model.

3.1. PhysioNet/CinC 2016

The PhysioNet/Computing in Cardiology (CinC) Challenge 2016 is the first attempt to address the lack of a large and open database of heart sound recordings: before it, the only two open-source datasets available counted together for less than 200 PCG signals making their use worthless to develop classification models [5, 30]. It is made by assembling databases collected from seven different research groups worldwide over more than ten years, in clinical and nonclinical environments. Due to its composite nature, signals from distinct sources differ in several aspects, such as the type of stethoscope, sampling rate, recording positions, number of recordings per individual, length of recorded signals, subjects' cardiac conditions, and methods for obtaining diagnoses. In particular, the 3M Littmann and the Welch Allyn Meditron electronic stethoscopes were used for the data acquisition of three and two databases respectively, whereas, in the other cases, the type of sensor is either unknown, prototypes, or from different manufacturers. In most cases, the sampling frequency chosen was either $4k$Hz or $8k$Hz, except for one database in which it was set at $44.1k$Hz. The number of recordings per subject, the sensor placement, and the signals' length differ both between and within databases, with recording duration ranging from several seconds to a few minutes, and stethoscope locations varying from one to multiple chest positions that do not always correspond to the standard auscultation ones. The database includes heart sounds from both healthy subjects and pathological patients, mainly affected by heart valve diseases (such as aortic stenosis or mitral regurgitation) and coronary artery disease. In most cases, the actual diagnoses were established by echocardiographic examination, although sometimes they were determined by clear heart murmurs or a cardiologist's diagnosis that is not further specified. In addition, demographic data, such as age and gender, were sometimes provided. Since the data were acquired in real-world, uncontrolled scenarios, many recordings are corrupted by noise, including sensor motion and speech [5].
The challenging training set is publicly available in PhysioNet repository [31]: the dataset is clearly imbalanced, consisting of a total of 3,153 heart sound signals from 764 individuals, with 2,488 recordings defined as normal and the remaining 665 obtained from pathological patients. In order to standardize the realized data, all the signals have been resampled at $2k$Hz and provided in .hea and .wav formats [30].

3.1.1. MITHSDB

The Massachusetts Institute of Technology Heart Sounds Database (MITHSDB) is one of the databases that contribute in composing PhysioNet/CinC. To the best of our knowledge, this is currently the only publicly available bimodal dataset that contains data from both healthy and pathological individuals. It is composed of 409 PCG recordings acquired using a Welch Allyn Meditron electronic stethoscope, of which 405 are coupled by a simultaneous single-lead ECG recording, obtained from 121 subjects.

Among the bimodal signals, 117 represent the normal control group, while the remaining 288 are labeled as pathological, having been collected from patients with either mitral valve prolapse, benign murmurs, aortic disease, or other miscellaneous pathological conditions. In this case, the diagnoses were verified based on the echocardiogram analysis [5].
Given that this is the only available bimodal dataset for disease classification task, it has attracted significant interest. Specifically, since deep neural networks need large amounts of data, P. Li et al. generated an augmented version of the dataset using a sliding window strategy. First of all, 17 visually noisy recordings were eliminated manually. The remaining 388 signals were first divided into training and validation datasets to generate mutually exclusive sets. At this stage, the data were expanded by segmenting the raw signals with a fixed window of $8s$, using a window stride of size $8s$ and $3s$ for abnormal and normal recordings respectively, in order to achieve a balance between the two classes. Then the PCG signals were resampled to $1k$Hz, whereas the frequency rate for ECGs was left unchanged at $2k$Hz. In total, this augmented version of the MITHSDB contains 1,975 recordings, 1,009 from healthy subjects and 966 from pathological ones, each with a fixed duration of $8s$ [3].
This expanded dataset will be leveraged in this work and can be downloaded on Zenodo [32].

3.2. GitHub Dataset

In 2018, an open-source heart sound database was released on GitHub. It contains 1,000 audio files from various sources, such as books and over 40 websites, each consisting of three-period heart sound signals and completely free of noise. The recordings are divided into 5 perfectly balanced groups, with 200 audio clips per category, corresponding to different cardiac valve conditions: normal, aortic stenosis, mitral regurgitation, mitral stenosis, and mitral valve prolapse. The recordings were sampled at $8k$Hz and are released in .wav format [13, 33].

3.3. CirCor DigiScope

The CirCor DigiScope database was collected during two screening campaigns performed in Northeast Brazil from a pediatric population (under 22 years old), with 70% of it publicly released for the George B. Moody PhysioNet Challenge 2022 [34]. The majority of the participants were children and infants, with most joining the study without a formal indication, others for follow-up on previously diagnosed heart conditions, and a smaller portion to monitor the progression of existing murmurs. The most common diagnoses included simple congenital cardiopathy and acquired cardiopathy, with some cases also diagnosed with complex congenital cardiopathy. In the database, the presence or absence of a pathological condition in the subjects was determined by a pediatric cardiologist's comprehensive assessment, which included clinical history, physical examination, and/or echocardiogram, but without having access to the signals recorded. The heart sounds were recorded primarily from one to four standard auscultation points (pulmonary, aortic, mitral, tricuspid) using a Littmann 3200 stethoscope equipped with the DigiScope Collector technology. Since the acquisitions were performed in a real clinical setting, the signals may have been corrupted by various noisy sources, such as stethoscope rubbing or background crying. The heart sound recordings were sampled at $4k$Hz, normalized within the [-1, 1] range, and provided in .hea and .wav formats. Additional demographic information, including age, gender, weight, and height, is provided for almost all the subjects [35]. Even if data quality assessment was performed, removing inconsistent and outliers data, this dataset appears to be noisy: researchers assert that the baseline accuracy is approximately equal to 55 % [36].
The public dataset consists of 3,163 recordings from 942 subjects, with 1,632 from individuals with normal heart conditions and 1,531 from those with pathological ones: it is provided online on PhysioNet [31].

4. Materials and Methods

This chapter outlines the experimental methodology, starting with a comprehensive literature review to establish the genesis of the study's rationale. Moreover, we describe the data preparation process and model configurations for both unimodal and bimodal analyses, explaining how transfer learning was applied to try to overcome data collection issues.

4.1. Comparative Literature Analysis

Analyzing the related works described in Section 2, key aspects have emerged that warrant further investigation to better understand the rationale behind the experiments conducted in this work and to derive fair results for comparison. Most of these studies have shown that ECG-only models tend to outperform PCG-only models when using the MITHSDB bimodal data in single-modality settings. This highlights the need for further efforts to enhance the interpretability of PCGs in order to improve the performance of bimodal models. Additionally, as outlined many times, the main challenge in developing bimodal models is the lack of an adequate dataset, a problem that transfer learning has been applied to address. Using this approach, Vieira [29] pointed out that the fine-tuning setting failed to outperform the feature extraction strategy from a pre-trained model on ImageNet, likely due to the limitation of relying solely on the Physionet/CinC 2016 dataset for the heart sound branch. This suggests the potential benefits of incorporating additional PCG datasets in such scenarios. Furthermore, Koike et al. showed that using a deep learning model pre-trained directly on audio data yields more valuable representations than transferring the knowledge from image data. In particular, they compared the performance of several models pre-trained on ImageNet, such as VGG-16, with that achieved using the Large-Scale Pretrained Audio Neural Networks (PANNs) [37] pre-trained on AudioSet, which had previously demonstrated strong generalization capabilities in many audio pattern recognition tasks. Using the Physionet/CinC 2016 PCG dataset, Koike et al. obtained the highest unweighted average recall in classifying normal versus abnormal heart sounds by leveraging the PANNs-CNN14 model to extract higher representations from the inputs [38]. Based on these findings, the experiments conducted in this study aim to explore a wiser transfer learning approach by fine-tuning the PANNs-CNN14 model, pre-trained on AudioSet, with additional unimodal PCG data.
Table 1 outlines the principal characteristics of all the bimodal studies described in Section 2. Notably, four of these studies rely on private databases rather than the MITHSDB, and the diseases predicted do not always align with the labels presented in MITHSDB. As a result, they are unsuitable for direct performance comparison with our solution. For the remaining studies, the best model performance are reported, particularly the AUROC (Area Under the Receiver Operating Characteristic Curve) when available, as well as the accuracy: the AUROC is independent of dataset imbalance, making it a more reliable metric compared to accuracy which is, in contrast, highly affected by different class distributions. For instance, J. Li et al. [21] used the MITHSDB which has an imbalanced class distribution, as well as the subset of MITHSDB extracted by Singh et al. in [20] that includes 240 abnormal and 102 normal signals. In this latter case, the model achieved a high accuracy of 93.1%, but this is largely due to its tendency to predict the majority class (abnormal). In addition to the dataset's class imbalance, in [25] the confusion matrix and the corresponding performance reported were based on the training set, raising concerns about overfitting. Another factor that raises questions about the performance reported by Morshed et al. [26] is that the test set performance is evaluated using only 10% of the data, but when the test set size increased to 30%, the performance significantly dropped, with accuracy falling to 90.7% and AUROC to 96%, suggesting a poor generalization ability of the model.

4.2. Experimental Pipeline

4.2.1. Data Preparation

The raw signals, both bimodal data from the MITHSDB and unimodal PCG recordings from unimodal datasets, undergo initial pre-processing steps such as filtering, downsampling, and normalization. The

Table 1
Summary of studies that combined ECG and PCG signals for classifying pathological heart conditions. AUROC = Area Under the Receiver Operating Characteristic Curve, BiLSTM = Bidirectional Long Short Term Memory, CAD = Coronary Artery Disease, CNNs = Convolutional Neural Networks, HD = Heart Disease, LSTM = Long Short Term Memory, SVM = Support Vector Machine.

Study	Outcome	Dataset	Approach used	Performance
H. Li et al. [18]	CAD/non-CAD	Augmented dataset from 195 subjects	Manually features extraction and deep learning	Accuracy = 95.6%
Chakir et al. [19]	Normal/Abnormal	100 subjects of MITHSDB	SVM	Accuracy = 92.5% AUROC = 95.1%
Singh et al. [20]	Normal/Abnormal	342 subjects of MITHSDB	SVM	Accuracy = 93.1%
J. Li et al. [21]	Normal/Abnormal	MITHSDB	SVM	Accuracy = 86.4%
EL-Bouridy and EL-Batouty [22]	Normal/Abnormal	Integrated cardiograph scanned image of 12-lead ECGs and 5-probe PCGs	Signals' decompositions and Neural Network	Accuracy = 96.8%
Jyothi and Pradeepini [23]	Five classes: normal, arrhythmias, mitral valve prolapse, ischemic HD, valvular HD	335 subjects	Signals' decompositions and Neural Network	Accuracy = 96.1%
P. Li et al. [3]	Normal/Abnormal	Augmented version of MITHSDB (3.1.1)	CNNs and LSTM	Accuracy = 87.3% AUROC = 93.6%
H. Li et al. [24]	CAD/non-CAD	Augmented dataset from 195 subjects	1-D and 2-D CNNs	Accuracy = 96.5%
J. Li et al. [25]	Normal/Abnormal	MITHSDB	BiLSTM-GoogLeNet	Accuracy = 96.1%
Morshed et al. [26]	Normal/Abnormal	Augmentation of MITHSDB	1-D CNNs	Accuracy = 95.1% AUROC = 99%
Hettiarachchi et al. [27]	Normal/Abnormal	Augmentation of MITHSDB	Transfer learning	Accuracy = 87.7% AUROC = 93.8%
Vieira [29]	Normal/Abnormal	Augmentation of MITHSDB	Transfer learning from ImageNet	Accuracy = 82.8% AUROC = 91.3%

most clinically relevant ECG components primarily occupy a frequency band below 50 Hz, whereas the frequency content of PCG typically falls within the 20-200 Hz range [39, 40]. Therefore, to reduce noise and simultaneously eliminate power-line interference in ECGs, a second-order low-pass Butterworth filter is applied to both signals, with cutoff frequencies set to 48 Hz for ECG and 200 Hz for PCG. The sufficiency of a low-pass filter for PCGs is attributed to the fact that the augmented MITHSDB provided recordings decomposed into four separate frequency bands ranging from 25 to 400 Hz, which were summed together before filtering. After that, the signals are downsampled to 100 Hz and 500 Hz respectively, and then normalized as follows:

$$x_{norm} = \frac{x - x_{mean}}{x_{max}}$$

To reduce the influence of spurious peaks caused by noise, the maximum value used for normalization is determined as the median of the maximum values detected within intervals obtained using a sliding window approach with a window length of 1 second, ensuring that each window almost always contains at least one heartbeat. By doing so, all the signals have zero mean and range almost in [-1, 1] interval. These steps are applied regardless of the source dataset. For the PCG recordings coming from unimodal datasets, additional transformations are applied. In particular, to align these signals with the ones within the augmented version of the MITHSDB used, recordings shorter than 8 seconds are replicated until the desired duration is reached, while recordings longer than 8 seconds are cropped. In the latter case, cropping is generally performed in the center, however, if the length of a recording exceeds 16 seconds (plus a 1-second margin of error, 0.5 seconds on each side), the signal is divided into multiple non-overlapping segments. Doing so, the number of PCGs in PhysioNet/CinC 2016 (excluding the signals from MITHSDB) increases from 2,744 to 5,857, the dimension of CirCor DigiScope grows from 3,163 to 7,076, while the size of the GitHub-dataset remains unchanged. Furthermore, before the

normalization step but after extending the recordings, Gaussian white noise is added to the signals from the GitHub-dataset to make the PCGs more similar to those in the other datasets.
Since the MITHSDB is already divided into two distinct datasets with an 80/20 split, the training and validation sets are created by further splitting the 80% portion into an 80/20 ratio, ensuring that each patient is assigned to only one group, keeping the 20% portion for the test set. As a result, the training set contains 1,243 recordings of both ECG and PCG, while the validation and test sets consist of 337 and 395 bimodal signals, respectively. On the other hand, each unimodal dataset is split into two separate groups with an 80/20 ratio, stratified by label before the extension step, and then combined. This results in an unimodal training set of 11,111 recordings, of which 7,124 are from subjects with normal heart conditions, and a validation PCG set consisting of a total of 2,823 signals (1,822 normal).

4.2.2. Model setting and configuration

As a first step toward the proposed bi-modal analysis system, one-dimensional Convolutional Neural Networks (1D-CNNs) based on either ECG or PCG signals are developed to classify normal and abnormal heart conditions using the unimodal portion's data from the augmented version of MITHSDB. Since abnormal heart conditions like valvular diseases often present with localized and specific waveform features, 1D-CNNs are a strategic choice due to their high ability to capture local and hierarchical spatial patterns through their receptive fields. In this way, the model can efficiently learn specific waveform features that indicate abnormalities, without the need for extensive modeling of long-term dependencies that other architectures, such as LSTMs, would focus on. The unimodal network consists of multiple convolutional blocks, each containing a 1-D convolutional layer, batch normalization, ReLU activation function, and a pooling layer with a stride of 2. The convolutional portion is followed by global average pooling and two fully connected layers (64 and 1 neurons respectively) with batch normalizations and non-linear transformations (ReLU and Sigmoid). Table 2 summarizes the different configurations of convolutional blocks, including the number of feature maps and kernel sizes associated, as well as the pooling types and whether dropout with a neglect probability of 0.2 is applied that are tested to optimize the model's performance; information about the models' complexity, in term of number of trainable parameters, is also provided.
The next step involves leveraging all the unimodal PCG datasets to perform transfer learning, using the PANNs-CNN14 as base model. The parameters of the initial layers, which are responsible for generating the spectrogram and its log-mel transformation, are modified to handle the fixed-size 8-second signals. Meanwhile, the weights obtained from training the PANNs-CNN14 model on AudioSet [37] are loaded into the remaining part of the network. The final fully connected layer, originally designed for multiclass classification across 527 classes in AudioSet, was replaced with a multilayer perceptron comprising four fully connected layers with 512, 128, 32, and 1 neurons, respectively. Each of these layers is followed by batch normalizations and ReLU transformations, except for the final layer, where a Sigmoid function is applied to obtain the abnormal class's probability. The entire network is either fully fine-tuned or trained with its first convolutional blocks frozen using all the combinations of the unimodal datasets. For instance, a total of seven dataset combinations: with all datasets assembled, including only GitHub-dataset and excluding it, including only Physionet/CinC and excluding it, and including only CirCor DigiScope and excluding it.
Finally, the multimodal model is created using a late fusion strategy. The ECG and PCG embeddings, obtained from the unimodal models (excluding the fully connected head), are concatenated and then processed by a multi-layer perceptron to produce the classification. We evaluate both the network with the simple convolutional PCG-only model and the architecture where the PCG branch is managed by the PANNs-CNN14 model pre-trained with all the publicly available heart sound datasets.
Figure 2 shows all the different strategies implemented.

Training Settings In our experiments, since we use a sigmoid activation function for the final layer, the loss function with which all the models are trained is the binary cross-entropy (BCE) loss, defined

Table 2
Different unimodal network configurations. In the configuration with both pooling types, the first $\left\lfloor \frac{\text{\# conv layers}}{2} \right\rfloor + 1$ pooling layers implement average pooling, while the remaining layers utilize max pooling.

# conv layers	Features maps	Kernel sizes	Pooling types	P dropout	# params
5	16, 32, 64, 128, 256	7, 5, 5, 3, 3	all avg/all max/both	0/0.2	153923
6	16, 32, 64, 128, 256, 256	7, 7, 5, 5, 3, 3	all avg/all max/both	0/0.2	368707
7	16, 32, 64, 64, 128, 256, 256	7, 7, 5, 5, 3, 3, 3	all avg/all max/both	0/0.2	372995
8	16, 32, 64, 64, 128, 128, 256, 256	7, 7, 5, 5, 5, 3, 3, 3	all avg/all max/both	0/0.2	438915

as:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right]$$

were N is the batch size, $\hat{y}_i$ is the predicted probability of the abnormal class and y_i is either 1 or 0, indicating whether the true label of the i-th input is abnormal or normal, respectively.
The batch size is set to 32 when using the augmented version of MITHSDB since the datasets are still small. However, the batch size is increased to 256 during the fine-tuning step with the unimodal PCG datasets to improve training stability. Moreover, an Adam optimizer with a learning rate of 0.01 is used for all the trainings, and the maximum number of epochs is set to 100. However, an early stopping technique with a patience of 15 and a tolerance of 0.001 is implemented, considering accuracy when the class distribution is balanced (i.e. in the MITHSDB) and AUROC in the imbalanced case.

5. Results

Below, we present the results of our experiments, reporting the test performance achieved by the best unimodal and bimodal model configurations. We place particular emphasis on evaluating the effectiveness of the transfer learning approach, examining its impact on the PCG interpretation and its influences on the overall performance of the bimodal configuration for CVD detection.

5.1. One-Dimensional Convolutional Neural Networks (1D-CNNs)

Using MITHSDB in a single-modality setting, all the different configurations reported in Table 2 are implemented for training the 1D-CNN PCG-only model, whereas, due to the input size and the dimensionality reduction across the layers, the analogous ECG-only model is trained with up to 7 convolutional layers. The best unimodal models are selected based on both AUROC and accuracy values on the test set. Concerning the ECG, using AUROC as a selection criterion, the best model configuration results in six 1D-convolutional blocks and average pooling for all layers, whereas, the best accuracy is achieved with five 1D-convolutional blocks implementing both max and average pooling. About the PCG, both maximum AUROC and accuracy are obtained using eight 1D-convolutional blocks with only average pooling. Figure 3 shows their confusion matrices on the test set, while all the performances are summarized in Table 3 where the metric used to choose the best model is highlighted.

5.2. Transfer learning

The weights from the PANNs-CNN14 model that achieves the best mean average precision on AudioSet released in Zenodo [41] are loaded into the transfer learning model's base (excluding the modified layers

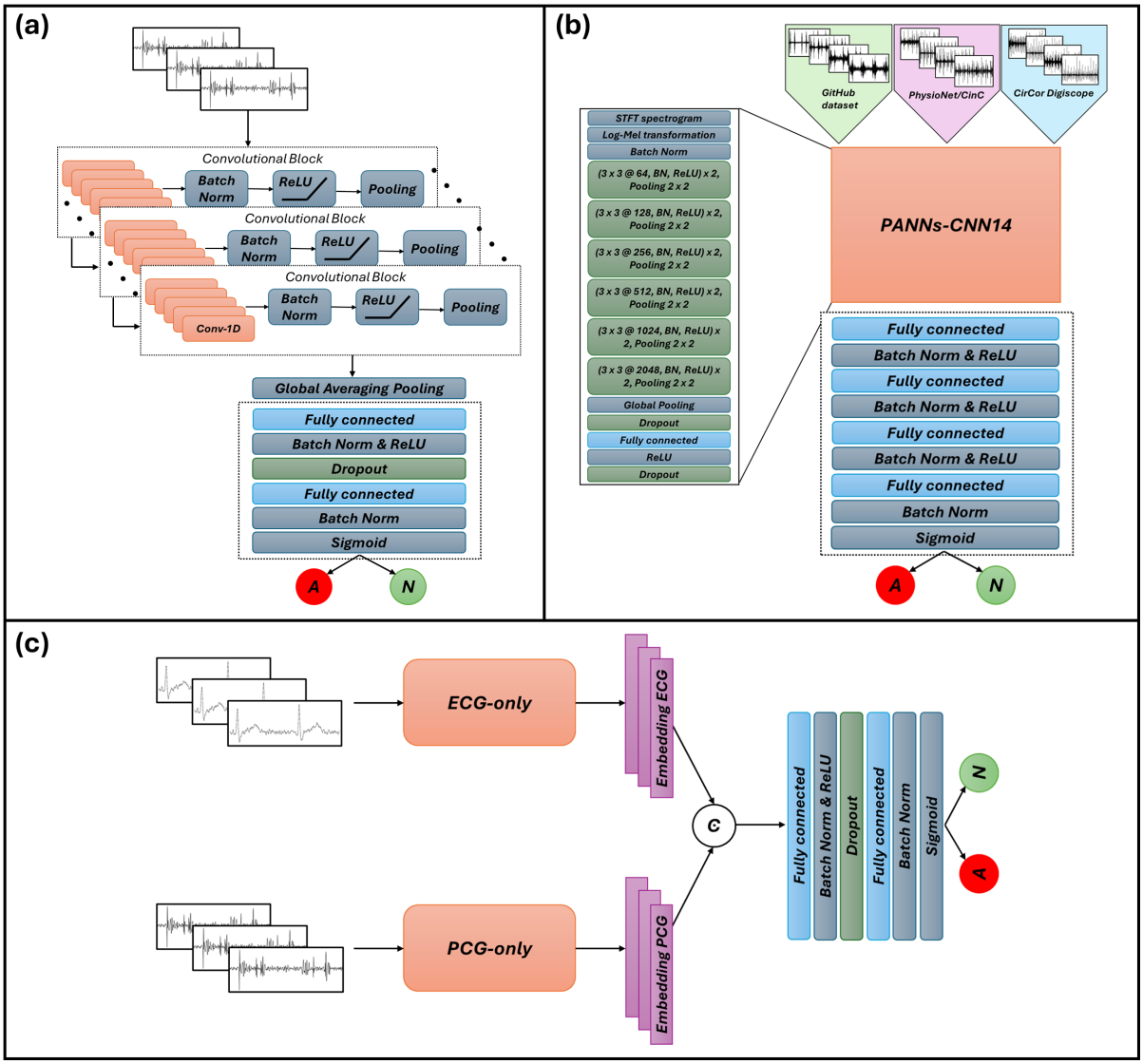

Figure 2: Graphical representation of the different architectures developed to classify between normal (*N*) and abnormal (*A*) heart conditions. **(a)**: Unimodal model consists of multiple 1D-convolutional blocks for either PCG or ECG signals. **(b)**: Transfer learning from PANNs-CNN14 using different combinations of unimodal PCG datasets for the fine-tuning process. After the symbol *@* is indicated the number of feature maps, while *BN* stands for batch normalization. **(c)**: Late fusion bimodal architecture using the embedding from the best ECG-only and PCG-only model.

for the log-mel extraction). Both a complete fine-tuning approach, where all layers in the base model are set as trainable, and a partial freezing approach, where the first 1, 2, or 3 out of the 6 convolutional blocks are frozen, are implemented. Additionally, all possible combinations of the PCG datasets are used for fine-tuning the entire network. Subsequently, the fine-tuned model on the unimodal data is used to evaluate performance on the MITHSDB dataset. In this phase, the first 4 or 5 convolutional blocks of the base model are kept frozen, while the remaining layers are further fine-tuned using the MITHSDB training data. The best model, which achieves both the maximum accuracy and AUROC on the test set, results in a first fine-tuning (keeping the first two convolutional blocks frozen) using all the PCG recordings available excluding those coming from the GitHub-dataset, and then further fine-tuned with MITHSDB training data with the first four convolutional blocks frozen. All test performance and confusion matrix are also reported respectively in Table 3 (last line) and Figure 3 (bottom right).

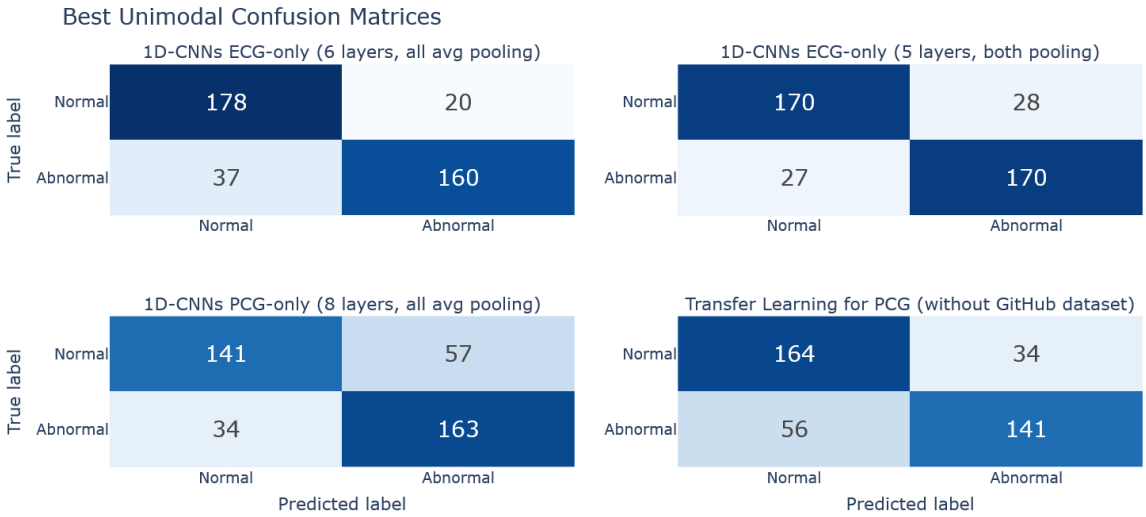

Figure 3: Confusion matrices of the best unimodal models on the test set of MITHSDB.

Table 3
Summary of the best unimodal performance on the test set extracted from the augmented version of MITHSDB. For each model, the metric value used to select the best one is highlighted.

Configuration	Accuracy	AUROC	Precision	Recall	Specificity	F1
1D-CNNs ECG-only; 5 layers, both pooling types	**86.1**%	92.6%	85.9%	86.3%	85.9%	86.1%
1D-CNNs ECG-only; 6 layers, all avg pooling type	85.6%	**93.1**%	88.9%	81.2%	89.9%	84.9%
1D-CNNs PCG-only; 8 layers, both pooling types, dropout	**77**%	**83**%	74.1%	82.7%	71.2%	78.2%
Transfer learning PCG	**77.2**%	**85.7**%	80.6%	71.6%	82.8%	75.8%

Table 4
Summary of the best bimodal performance on the test set extracted from the augmented version of MITHSDB, with or without implementing transfer learning into the PCG branch. For each model, the metric value used to select the best one is highlighted.

ECG model	PCG model	Accuracy	AUROC	Precision	Recall	Specificity	F1
1D-CNNs 6 layers, all avg pooling type; frozen	1D-CNNs 8 layers, both pooling types, dropout; frozen	**90.6** %	**96.4**%	88.1%	93.9%	87.4%	90.9%
1D-CNNs 6 layers, all avg pooling type; frozen	Transfer learning unimodal PCG model; fronzen	**90.4**%	**95.8**%	90%	91.4%	89.4%	90.5%

5.3. Bimodal

Each of the best unimodal models obtained above, with their "head" (identified with a dashed rectangle in Figure 2) removed, is then incorporated into the bimodal network. In this setup, both strategies of fine-tuning all the bimodal network and keeping one or both unimodal branches frozen to extract embeddings are tested. The best configurations, selected based on both test accuracy and AUROC, with or without the implementation of transfer learning into the PCG branch, see the six 1D-convolutional block configuration for the ECG branch as well as both branches frozen during the embedding extraction. The results of the bimodal analysis are summarized in Table 4, while in Figure 4 the confusion matrices on the MITHSDB test set are reported.

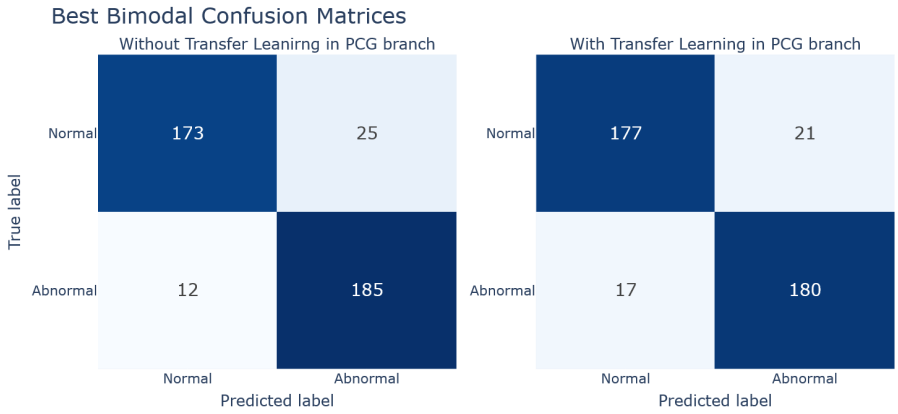

Figure 4: Confusion matrices of the best bimodal models on the MITHSDB test set with and without using the transfer learning approach to extract highly representative features from heart sound recordings.

6. Discussion

Concerning the unimodal approach, our findings confirm the results of most prior works. Specifically, when using the MITHSDB in a single-modality setting, the best models developed using only ECG signals outperform those trained on the corresponding PCG recordings. Additionally, despite fine-tuning the PANNs-CNN14 model with a larger and more diverse set of heart sounds, it doesn't improve the PCG interpretability as much as we expected compared to the simpler 1D-CNNs model. This result suggests that even a more reasonable transfer learning approach can not completely overcome the limitations posed by the lack of bimodal data, emphasizing the need for additional research and data collection in this area. Moreover, from the confusion matrices reported in Figure 3, we notice that the transfer learning approach tends to predict normal labels, leading to a high number of false negatives (i.e., misclassifying pathological heart sounds as normal). In contrast, the model based on 1D-CNNs is more inclined to assign abnormal labels, which is preferable in a CVD screening setting. This suggests that an ensemble model could improve the ability to distinguish between normal and abnormal heart sounds. This reasoning, along with the analysis of the confusion matrices for the unimodal ECG and PCG models, may also explain why the limited improvements obtained using transfer learning, are not reflected in the bimodal setting. For instance, both the 1D-CNNs ECG-only with six convolutional blocks and the PANNs-CNN14 fine-tuned models are more likely to predict normal labels, resulting in a higher amount of false negatives, whereas the 1D-CNNs PCG-only model is more effective at identifying the abnormal class. Therefore, by combining the strengths of both the ECG and PCG branches, the bimodal model is able to take advantage from each modality, significantly improving overall performance and further reducing the false-negative rate. As highlighted in Section 4.1, a direct comparison with existing bimodal approaches is not always feasible for several reasons. However, a fair comparison with previous works can be made when the classification task is consistent, and key information that affects performance, such as dimensionality, is provided. In such cases, we can rely on the AUROC metric, as it retains clinical relevance even in unbalanced datasets. Based on these considerations, as reported in the comparison Table 5, our best bimodal configuration achieves a higher AUROC on unseen data compared to the other existing bimodal approaches. This result suggests that valuable information can be extracted directly from the temporal characteristics of these signals without the need for manual feature extraction or a 2D time-frequency representation. Additionally, we mitigate the overfitting issue that can arise when data is limited and the model has too many parameters, as may be the case for P. Li et al. [3], by reducing the model's complexity. Despite the model's strong performance, its generalizability remains limited due to reliance on a single bimodal dataset. To assess the model's adaptability in real-world scenarios, particularly across diverse patient populations, a more comprehensive data acquisition process is crucial. Another limitation of this work involves the absence of explainability techniques, avoiding a clearer understanding of which modality contributes most significantly to the predictions. Identifying the contributions of each modality would provide valuable

Table 5
Perfomance comparison with related works.

Author	AUROC	Accuracy	Recall	Specificity
Chakir et al. [19]	95.1%	92.5%	92.3%	92.9%
P. Li et al. [3]	93.6%	87.3%	90.3%	84.5%
Hettiarachchi et al. [27]	93.8%	87.7%	87.7%	87.5%
Vieira [29]	91.3%	82.8%	93.1%	57.1%
Our model	**96.4**%	**90.6**%	**93.9**%	**87.4**%

insights into the model's decision-making process, enhancing interpretability and supporting more reliable clinical applications. Addressing this limitation could be a key focus for future development.

7. Conclusions

In this work, we conducted an in-depth analysis of previous research regarding the use of ECG and PCG signals in developing ML models for early CVD detection. We explored a promising transfer learning approach by leveraging publicly available unimodal PCG datasets to address the challenge of insufficient bimodal ECG-PCG data. However, our findings indicate that this approach does not fully answer the limitations posed by the lack of enough bimodal datasets, highlighting the need for further efforts in this area. Nevertheless, our results confirm the effectiveness of a multimodal approach, demonstrating that combining the complementary insights from ECG and PCG signals can significantly enhance CVD detection compared to single-modality approaches. Our proposed bimodal model, built on 1D-CNN architectures and employing a late fusion strategy, outperforms existing methods, achieving an AUROC of 96.4% in abnormal heart condition detection, and demonstrating its potential for integration into large-scale CVD screening systems.

References

[1] World Health Organization, Cardiovascular diseases, URL: https://www.who.int/health-topics/cardiovascular-diseases, 2024. Accessed: 2024-08-20.

[2] R. C. Joshi, J. S. Khan, V. K. Pathak, M. K. Dutta, Ai-cardiocare: Artificial intelligence based device for cardiac health monitoring, IEEE Transactions on Human-Machine Systems 52 (2022) 1292–1302. doi:10.1109/THMS.2022.3211460.

[3] P. Li, Y. Hu, Z.-P. Liu, Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods, Biomedical Signal Processing and Control 66 (2021) 102474. doi:10.1016/j.bspc.2021.102474.

[4] S. Singh, R. Chaudhary, K. P. Bliden, U. S. Tantry, P. A. Gurbel, S. Visweswaran, M. E. Harinstein, Meta-analysis of the performance of ai-driven ecg interpretation in the diagnosis of valvular heart diseases, The American Journal of Cardiology 213 (2024) 126–131. doi:10.1016/j.amjcard.2023.12.015.

[5] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. W. Johnson, Z. Syed, S. E. Schmidt, C. D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M. R. Samieinasab, R. Sameni, R. G. Mark, G. D. Clifford, An open access database for the evaluation of heart sound algorithms, Physiological Measurement 37 (2016) 2181. doi:10.1088/0967-3334/37/12/2181.

[6] M. Boulares, R. Alotaibi, A. AlMansour, A. Barnawi, Cardiovascular disease recognition based on heartbeat segmentation and selection process, International Journal of Environmental Research and Public Health 18 (2021). doi:10.3390/ijerph182010952.

[7] S. Guo, B. Zhang, Y. Feng, Y. Wang, G. Tse, T. Liu, K.-Y. Chen, Impact of automatic acquisition of key clinical information on the accuracy of electrocardiogram interpretation: a cross-sectional study, BMC Medical Education 23 (2023) 936. doi:10.1186/s12909-023-04907-9.

[8] F. Wang, T. Syeda-Mahmood, D. Beymer, Finding disease similarity by combining ecg with heart auscultation sound, in: 2007 Computers in Cardiology, 2007, pp. 261–264. doi:10.1109/CIC.2007.4745471.

[9] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, U. R. Acharya, Classification of myocardial infarction with multi-lead ecg signals and deep cnn, Pattern Recognition Letters 122 (2019) 23–30. doi:10.1016/j.patrec.2019.02.016.

[10] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. M. Jr., T. B. Schön, A. L. P. Ribeiro, Automatic diagnosis of the 12-lead ecg using a deep neural network, Nature Communications 11 (2020) 1760. doi:10.1038/s41467-020-15432-4.

[11] J. S. Haimovich, N. Diamant, S. Khurshid, P. Di Achille, C. Reeder, S. Friedman, P. Singh, W. Spurlock, P. T. Ellinor, A. Philippakis, P. Batra, J. E. Ho, S. A. Lubitz, Artificial intelligence–enabled classification of hypertrophic heart diseases using electrocardiograms, Cardiovascular Digital Health Journal 4 (2023) 48–59. doi:10.1016/j.cvdhj.2023.03.001.

[12] F. A. Khan, A. Abid, M. S. Khan, Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features, Physiological Measurement 41 (2020) 055006. doi:10.1088/1361-6579/ab8770.

[13] Y. Al-Issa, A. M. Alqudah, A lightweight hybrid deep learning system for cardiac valvular disease classification, Scientific Reports 12 (2022) 14297. doi:10.1038/s41598-022-18293-7.

[14] M. T. Nguyen, W. W. Lin, J. H. Huang, Heart sound classification using deep learning techniques based on log-mel spectrogram, Circuits, Systems, and Signal Processing 42 (2023) 344–360. doi:10.1007/s00034-022-02124-1.

[15] A. Kazemnejad, S. Karimi, P. Gordany, G. D. Clifford, R. Sameni, An open-access simultaneous electrocardiogram and phonocardiogram database, Physiological Measurement 45 (2024). doi:10.1088/1361-6579/ad43af.

[16] Z. Ren, Y. Chang, T. T. Nguyen, Y. Tan, K. Qian, B. W. Schuller, A comprehensive survey on heart sound analysis in the deep learning era, IEEE Computational Intelligence Magazine 19 (2024) 42–57. doi:10.1109/MCI.2024.3401309.

[17] S. Ismail, B. Ismail, I. Siddiqi, U. Akram, Pcg classification through spectrogram using transfer learning, Biomedical Signal Processing and Control 79 (2023) 104075. doi:10.1016/j.bspc.2022.104075.

[18] H. Li, X. Wang, C. Liu, Y. Wang, P. Li, H. Tang, L. Yao, H. Zhang, Dual-input neural network integrating feature extraction and deep learning for coronary artery disease detection using electrocardiogram and phonocardiogram, IEEE Access 7 (2019) 146457–146469. doi:10.1109/ACCESS.2019.2943197.

[19] F. Chakir, A. Jilbab, C. Nacir, A. Hammouch, Recognition of cardiac abnormalities from synchronized ecg and pcg signals, Physical and Engineering Sciences in Medicine 43 (2020) 673–677. doi:10.1007/s13246-020-00875-2.

[20] S. Ajitkumar Singh, S. Ashinikumar Singh, N. Dinita Devi, S. Majumder, Heart abnormality classification using pcg and ecg recordings, Computación y Sistemas 25 (2021) 381–391. doi:10.13053/cys-25-2-3447.

[21] J. Li, L. Ke, Q. Du, X. Chen, X. Ding, Multi-modal cardiac function signals classification algorithm based on improved d-s evidence theory, Biomedical Signal Processing and Control 71 (2022) 103078. doi:10.1016/j.bspc.2021.103078.

[22] M. E. EL-Bouridy, A. S. EL-Batouty, An intelligent high accuracy hybrid identification for heart diseases diagnosis, in: 2021 International Telecommunications Conference (ITC-Egypt), 2021, pp. 1–5. doi:10.1109/ITC-Egypt52936.2021.9513892.

[23] P. Jyothi, G. Pradeepini, Heart disease detection system based on ecg and pcg signals with the aid of gkvdlnn classifier, Multimedia Tools and Applications 83 (2024) 30587–30612. doi:10.1007/s11042-023-16562-9.

[24] H. Li, X. Wang, C. Liu, P. Li, Y. Jiao, Integrating multi-domain deep features of electrocardiogram and phonocardiogram for coronary artery disease detection, Computers in Biology and Medicine

138 (2021) 104914. doi:10.1016/j.compbiomed.2021.104914.
[25] J. Li, L. Ke, Q. Du, X. Ding, X. Chen, Research on the classification of ecg and pcg signals based on bilstm-googlenet-ds, Applied Sciences 12 (2022). doi:10.3390/app122211762.
[26] M. Morshed, S. A. Fattah, A deep neural network for heart valve defect classification from synchronously recorded ecg and pcg, IEEE Sensors Letters 7 (2023) 1–4. doi:10.1109/LSENS.2023.3307053.
[27] R. Hettiarachchi, U. Haputhanthri, K. Herath, H. Kariyawasam, S. Munasinghe, K. Wickramasinghe, D. Samarasinghe, A. De Silva, C. U. S. Edussooriya, A novel transfer learning-based approach for screening pre-existing heart diseases using synchronized ecg signals and heart sounds, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021, pp. 1–5. doi:10.1109/ISCAS51556.2021.9401093.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.
[29] H. M. C. Vieira, Multimodal Deep Learning for Heart Sound and Electrocardiogram Classification, Master's thesis, Master's thesis, University of Porto, Porto, Portugal, 2023.
[30] 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: 2016 Computing in Cardiology Conference (CinC), 2016, pp. 609–612.
[31] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet, Circulation 101 (2000) e215–e220. doi:10.1161/01.CIR.101.23.e215.
[32] P. Li, Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods, 2020. doi:10.5281/zenodo.4263528.
[33] Yaseen, G.-Y. Son, S. Kwon, Classification of heart sound signal using multiple features, Applied Sciences 8 (2018). doi:10.3390/app8122344.
[34] J. Oliveira, F. Renna, P. Costa, M. Nogueira, A. C. Oliveira, A. Elola, C. Ferreira, A. Jorge, A. B. Rad, M. Reyna, R. Sameni, G. Clifford, M. Coimbra, The circor digiscope phonocardiogram dataset (version 1.0.3), PhysioNet, 2022. URL: https://doi.org/10.13026/tshs-mw03.
[35] J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares, A. Elola, A. B. Rad, R. Sameni, G. D. Clifford, M. T. Coimbra, The circor digiscope dataset: From murmur detection to murmur classification, IEEE Journal of Biomedical and Health Informatics 26 (2022) 2524–2535. doi:10.1109/JBHI.2021.3137048.
[36] G. Singh, A. Verma, L. Gupta, A. Mehta, V. Arora, An automated diagnosis model for classifying cardiac abnormality utilizing deep neural networks, Multimedia Tools and Applications 83 (2024) 39563–39599. doi:10.1007/s11042-023-16930-5.
[37] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894. doi:10.1109/TASLP.2020.3030497.
[38] T. Koike, K. Qian, Q. Kong, M. D. Plumbley, B. W. Schuller, Y. Yamamoto, Audio for audio is better? an investigation on transfer learning models for heart sound classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2020, pp. 74–77. doi:10.1109/EMBC44109.2020.9175450.
[39] L. G. Tereshchenko, M. E. Josephson, Frequency content and characteristics of ventricular conduction, Journal of Electrocardiology 48 (2015) 933–937. doi:10.1016/j.jelectrocard.2015.08.034.
[40] S.-Y. Lee, P.-W. Huang, J.-R. Chiou, C. Tsou, Y.-Y. Liao, J.-Y. Chen, Electrocardiogram and phonocardiogram monitoring system for cardiac auscultation, IEEE Transactions on Biomedical Circuits and Systems 13 (2019) 1471–1482. doi:10.1109/TBCAS.2019.2947694.
[41] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. Plumbley, PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition (Pretrained Models), 2020. doi:10.5281/zenodo.3987831.