





Article

Phonocardiogram (PCG) Murmur Detection Based on the Mean Teacher Method

Yi Luo ^{1,†}, Zuoming Fu ^{2,†} , Yantian Ding ³ , Xiaojian Chen ¹  and Kai Ding ^{4,*} 

¹ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21287, USA; yluo62@jh.edu (Y.L.); xchen279@jh.edu (X.C.)

² Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA; zf242@cornell.edu

³ Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21287, USA; yding72@jh.edu

⁴ Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD 21287, USA

* Correspondence: kai@jh.edu

† These authors contributed equally to this work.

Abstract: Cardiovascular diseases (CVDs) are among the primary causes of mortality globally, highlighting the critical need for early detection to mitigate their impact. Phonocardiograms (PCGs), which record heart sounds, are essential for the non-invasive assessment of cardiac function, enabling the early identification of abnormalities such as murmurs. Particularly in underprivileged regions with high birth rates, the absence of early diagnosis poses a significant public health challenge. In pediatric populations, the analysis of PCG signals is invaluable for detecting abnormal sound waves indicative of congenital and acquired heart diseases, such as septal defects and defective cardiac valves. In the PhysioNet 2022 challenge, the murmur score is a weighted accuracy metric that reflects detection accuracy based on clinical significance. In our research, we proposed a mean teacher method tailored for murmur detection, making full use of the Phyionet2022 and Phyionet2016 PCG datasets, achieving the SOTA (State of Art) performance with a murmur score of 0.82 and an AUC score of 0.90, providing an accessible and high accuracy non-invasive early stage CVD assessment tool, especially for low and middle-income countries (LMICs).

Keywords: cardiovascular disease (CVD); phonocardiogram (PCG); murmur detection; semi-supervised learning; mean teacher; non-invasive assessment; low and middle-income countries (LMICs)



Citation: Luo, Y.; Fu, Z.; Ding, Y.; Chen, X.; Ding, K. Phonocardiogram (PCG) Murmur Detection Based on the Mean Teacher Method. *Sensors* **2024**, *24*, 6646. <https://doi.org/10.3390/s24206646>

Academic Editor: Xuming Zhang

Received: 9 August 2024

Revised: 3 October 2024

Accepted: 3 October 2024

Published: 15 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The leading cause of mortality worldwide are CVDs [1]. Cardiac auscultation is the most accessible method for early-stage, non-invasive assessment of CVDs, particularly in LMICs. This technique involves the analysis of heart sounds, crucial indicators of heart function, and potential abnormalities. In the realm of digital health technology, PCGs elevate this basic assessment by recording these heart sounds electronically. PCGs capture the acoustic data of the heart's beats, which can be analyzed to detect irregular patterns such as murmurs, gallops, and rubs.

In PCG recordings, the first heart sound (S1) and the second heart sound (S2) are fundamental components. S1 is produced by the closing of the atrioventricular valves, signaling the beginning of systole. In contrast, S2 is generated by the closing of the semilunar valves, marking the end of systole and the start of diastole. Beyond these, heart murmurs detected in PCG signals provide essential clinical insights into heart malfunctions due to congenital and acquired diseases in pediatric populations [2,3]. Consequently, developing machine learning algorithms that can precisely detect murmurs from PCG signals presents a significant challenge.

One major challenge in applying deep learning techniques to cardiovascular signals is the lack of annotated data. Creating reliable diagnoses requires considerable time and the expertise of highly skilled professionals, making the process demanding and resource-intensive.

Although researchers have conducted lots of work to make full use of the labeled PCG data [4–6], a considerable portion of unlabeled data remain underutilized. Semi-supervised learning paradigms have demonstrated notable efficacy in leveraging unlabeled data alongside a limited subset of labeled data [7,8].

The PhysioNet 2022 [9] dataset comprises heart sound recordings collected from 1568 patients, most of whom were pediatric patients from rural Brazil. These recordings were gathered using an electronic stethoscope at four key auscultation points (aortic valve, pulmonic valve, tricuspid valve, and mitral valve). A total of 5272 recordings were made, with detailed annotations provided by cardiologists. The dataset includes demographic and clinical information such as age, sex, height, and weight, as well as clinical outcomes like heart murmurs and abnormal cardiac function. The recordings are labeled into three categories—Present, Absent, and Unknown for murmurs—providing a rich source of supervised learning data. However, only 60% of this dataset (2982 recordings) was released as the public portion for model training, which we utilized in our study.

In contrast, the PhysioNet 2016 dataset [10] was primarily collected for the classification of normal versus abnormal heart sounds and includes 4430 recordings from 1072 subjects. This dataset features recordings from a diverse population of both children and adults, gathered across various environments using different recording devices. While the primary objective was to differentiate between normal and abnormal heart function, the dataset lacks the detailed murmur annotations present in the PhysioNet 2022 dataset, making it well-suited as unlabeled data for our semi-supervised learning framework. The challenge's training set comprises 3153 heart sound recordings from 764 patients. Notably, the high-quality 2016-e subset, which has even been used for generative tasks [11], was included in our experiments and consists of 2054 PCG recordings.

In this study, we utilized the mean teacher model that leverages the labeled PCG data from Physionet 2022 in conjunction with the unlabeled PCG data in Physionet 2016, our method exhibits a comparable performance to fully supervised models. Additionally, by incorporating unlabeled data sourced from an external dataset alongside the labeled dataset, we achieve an enhanced model performance relative to utilizing labeled data alone. Our findings underscore the utility of integrating unlabeled data through a semi-supervised framework, thereby augmenting model performance and facilitating clinical decision-making by harnessing the wealth of readily available unlabeled data.

2. Related Work

Computer-aided classification in PCGs has captivated researchers for over 50 years, yet it continues to present challenges. Gerbarg et al. were pioneers in this field, using a threshold-based method to tackle the problem [12]. Early approaches in heart sound murmur detection involved enhancing acoustic signals through digital stethoscopes and basic electronic filtering techniques. Subsequently, techniques like the Fourier transform were employed to break down heart sounds into individual frequency components, facilitating the identification of abnormal frequencies [13,14]. Ahlström's work comprehensively summarized the physics, analysis, and processing of heart sounds [15]. More sophisticated approaches later integrated complex signal processing methodologies, such as spectral analysis and wavelet transformations [16–18]. Despite these advancements, traditional methods were limited by their dependency on high-quality acoustic signals and struggled in noisy clinical environments or with low-amplitude murmur sounds.

The rise of machine learning, especially deep learning, has transformed the landscape of murmur detection. The George B. Moody PhysioNet Challenge in 2016 [10] attracted many researchers to the field, leading to significant contributions in CVD assessment. Cristhian Potes et al. [19] achieved the highest score in the challenge, which combined the results from an AdaBoost algorithm and a convolutional neural network (CNN). The Ad-

aBoost algorithm was trained on 124 time–frequency features, while the CNN was trained on four frequency bands extracted from the PCG signals. Subsequent research utilized methods such as ResNet [20], Mel-frequency features [21], feature fusion methods [22], CNN-LSTM networks [23,24], and transfer learning [25] have achieved even higher performance.

The George B. Moody PhysioNet Challenge in 2022 [9] provided more detailed classifications for typical diagnoses, including heart murmurs. Hui Lu et al. achieved SOTA performance in murmur detection using a lightweight CNN with wide features and Mel-spectrogram [4], while LCNN, dual Bayesian ResNet, hidden semi-Markov model, and hierarchical multi-scale convolutional neural networks were also deployed to achieve competitive performance in the Physionet 2022 challenge [5,6,26,27]. Derivative works include the use of a new wav2vec model [28], multi-scale CNN with LSTM [29], and complex-valued networks [30]. Beyond the pursuit of higher classification performance, efforts have also been made to enhance model interpretability. For instance, M. Rohr et al. employed a multiple-instance learning framework to facilitate model interpretability [31], while G. Zhou et al. explored different data augmentation methods for PCG classification [32].

In the realm of semi-supervised methods, Bachman et al. [33] introduced the concept of “Pseudo-Ensembles” in 2014, which created an ensemble of models by perturbing the data and model parameters during training. This approach aims to enhance diversity in the learned representations. Further advancements were made by M Sajjadi et al. [34] in 2016, who introduced randomness into the training process to promote smoothness in the model’s predictions and increase robustness to variations in the input data [35]. Building upon this, Laine and Valpola proposed “Temporal Ensembling” in 2016, where the model’s predictions on unlabeled data were averaged over time to provide consistent pseudo-labels for regularization. These developments culminated in the introduction of the mean teacher concept by Tarvainen and Valpola in 2017 [8]. The mean teacher approach enhanced consistency regularization by encouraging the model to maintain stable predictions even when input data were slightly altered. This approach leverages both labeled and unlabeled data by comparing the student model’s current predictions with those of the teacher model, enabling simultaneous learning from both data types. This strategy greatly boosts the model’s ability to generalize and perform well on unseen data.

Despite the success of semi-supervised learning in various domains, its application in PCG murmur detection remains largely unexplored. The complexities of heart sound and the scarcity of labeled data pose significant challenges for traditional semi-supervised methods. However, integrating semi-supervised learning into PCG murmur detection could potentially alleviate the need for large labeled datasets and further improve classification performance in real-world clinical settings.

3. Method

Our method comprises two main components: the mean teacher model and a lightweight CNN architecture that serves as the base model.

The mean teacher method is a semi-supervised learning strategy that harnesses unlabeled data to significantly improve model generalization and robustness [8]. Compared to traditional semi-supervised learning approaches, the mean teacher approach employs a teacher model to enhance supervision, thereby reducing the generalization error and improving the model performance; the architecture of the mean teacher model is shown in Figure 1.

In the training process of the mean teacher method, student and teacher models both utilize the same architectural framework. During each iteration, a batch of labeled data is fed into the student model, generating predictions $p_{s,1}$. Unlabeled data are also input into the student model and subsequently into the teacher model, producing predictions $p_{s,2}$ and $p_{t,2}$, respectively. The classification loss between $p_{s,1}$ and the true labels, as well as the consistency loss between $p_{s,2}$ and $p_{t,2}$, are then calculated. These losses are weighted and combined to compute the total loss, which is used to optimize the student model, as indicated by the dotted line in Figure 1. Finally, the teacher model is updated through the

exponential moving average according to the parameters of the student model, completing one iteration. This process is repeated multiple times during a training cycle. As the iterations progress, the model's performance stabilizes, marking the completion of training. At the end of this process, the student model is employed for testing.

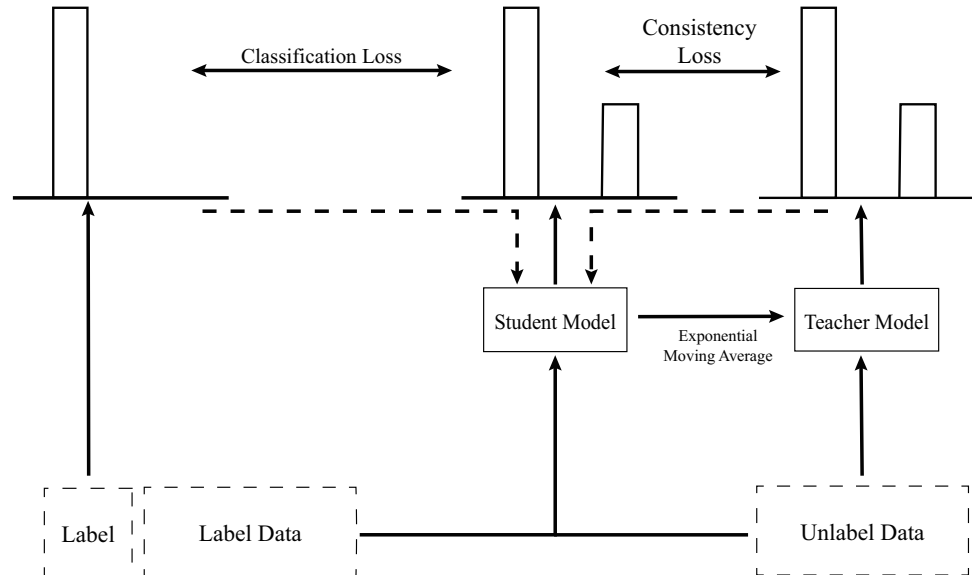


Figure 1. Architecture of Mean teacher model.

In the optimization process, the relative weighting of the classification and consistency losses dictates how supervised and unsupervised learning contribute to the model's performance. The mathematical expressions for these classification and consistency losses are outlined as follows.

$$\text{Classification Loss} = w_1 \sum_{i=1}^L y_i \log(p_{s;i}) \quad (1)$$

$$\text{Consistency Loss} = w_2 \sum_{i=1}^L (p_{t;i} - p_{s;i})^2 \quad (2)$$

In this project, weights w_1 and w_2 are assigned to the classification and consistency losses, respectively. The classification task involves three classes ($L = 3$). The label for class i is denoted as y_i , while $p_{s;i}$ and $p_{t;i}$ are the probabilities predicted by the student and teacher models for class i . The weight for the consistency loss is calculated using the function $e^{-r(1-x)^2} * (w_e - w_i) + w_i$, where $x = \frac{ep}{eps}$, with r representing the coefficient for the consistency loss weight. The classification loss weight is determined by a function given by $1 - \frac{ep-1}{eps}$, where ep represents the current epoch and eps is the total number of epochs. This function gradually decreases from 1 to 0 throughout the training period, allowing the model to initially converge quickly on labeled data and then progressively learn from unlabeled data.

The formula for Exponential Moving Average (EMA) is as follows:

$$\alpha_t = k\alpha_t + (1 - k)\alpha_s \quad (3)$$

In this context, α_t and α_s represent the trainable parameters of the teacher and student models, respectively. The smoothing coefficient k is set to 0.98 throughout the training process.

The following represents the foundational architecture of the teacher and student models, a lightweight CNN [4]; the architecture is shown in Figure 2. The lightweight CNN is the SOTA model in the Physionet 2022 challenge, outperforming other models due to its simple architecture, strong robustness, and generalization ability across dif-

ferent PCG signal data and populations. This makes it highly practical for use in PCG murmur detection.

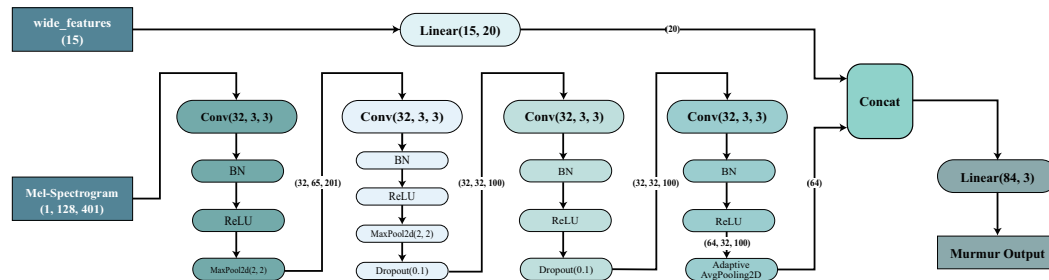


Figure 2. Architecture of the lightweight CNN model employed in mean teacher.

This lightweight convolutional neural network architecture is specifically designed for processing audio data, particularly for identifying cardiac murmurs. The primary input to the model is a Mel-spectrogram with dimensions of (1, 128, 401), where 128 represents the frequency dimension and 401 represents the time dimension. Additionally, the model receives an auxiliary input called “wide features”, which is a 15-dimensional feature vector. This vector is processed through a linear layer to transform it into a 20-dimensional feature.

The model architecture consists of four convolutional blocks followed by a linear layer. Each convolutional block includes batch normalization and a ReLU activation function to enhance audio feature extraction. The first and second convolutional blocks have max-pooling layers with a (2, 2) window, which reduces the spatial dimensions of the feature maps and lessens the computational burden. Additionally, the second convolutional block features a dropout layer with a 0.1 dropout rate to mitigate overfitting. Following the convolutional blocks, an adaptive average pooling layer is used to further condense the feature space before passing it to the linear layer.

Finally, the model combines 20-dimensional “wide features” processed through a linear layer with the 64-dimensional features derived from the Mel-spectrogram via the adaptive average pooling layer, resulting in an 84-dimensional feature vector. This vector is then passed through another linear layer, which maps it to one of three output categories: “Present”, “Absent”, or “Unknown”. The design of the entire model aims to optimize audio classification accuracy by learning the intrinsic features of the Mel-spectrogram and the directly inputted “wide features”. Additionally, the simplified network structure maintains computational efficiency.

4. Experiment

The development environment for our project was built using Python 3.9.0, Nvidia A100 GPUs, Pytorch 2.1.0, and CUDA 12.1.0. For the training process, we implemented the ADAM optimizer, setting the learning rate to 1e-3 and using a batch size of 16. The detailed explanation is shown in Table 1.

Table 1. Key parameter configuration.

Parameter Name	Parameter Value	Parameter Description
Optimizer	Adam	Neural Network optimizer
Learning Rate	1×10^{-3}	learning rate for optimizer
Batch Size	16	Number of samples per iteration.
EMA k	0.98	Smoothing coefficient in Exponential Moving Average.
Weight con init	0.7	Initial value for consistency loss in mean teacher.
Weight con end	1	Final value for consistency loss in mean teacher.

4.1. Dataset and Preprocessing

In our research, we primarily utilized data from the George B. Moody PhysioNet Challenge for the years 2016 and 2022. The 2022 dataset comprises heart sound recordings from 874 patients, totaling 2982 entries collected from various auscultation points. Additionally, this dataset includes annotated patient information such as age, sex, height, weight, and other relevant variables. The recordings are categorized into three main groups based on the presence of murmurs in the PCG signals: Present, Absent, and Unknown. These categories serve as the labeled data for the model training. The 2016 dataset, designed for normal/abnormal classification rather than murmur detection, includes the 2016-e datasets renowned for their high-quality 2054 PCG recordings, which we used as unlabeled data in our analysis.

In our project, we standardized the duration of audio recordings, which originally ranged from 4.8 s to 120 s, by setting a fixed duration of 10 s for each recording. During preprocessing, recordings longer than 10 s were randomly cropped, while shorter recordings were padded with zeros to achieve a consistent length. This preprocessing method was uniformly applied during both the training and validation phases.

For each recording, we extracted a Mel-spectrogram with 128 Mel bands, using a window size of 50 ms and a frame shift of 25 ms, covering a frequency range from 25 to 2000 Hz, as illustrated in Figure 3. In addition to spectrogram extraction, we computed wide features, which were then categorized and encoded into one-hot vectors. To further enrich the dataset, we extracted statistical features from the audio signals, including the zero-crossing rate, spectral centroids, and spectral bandwidth.

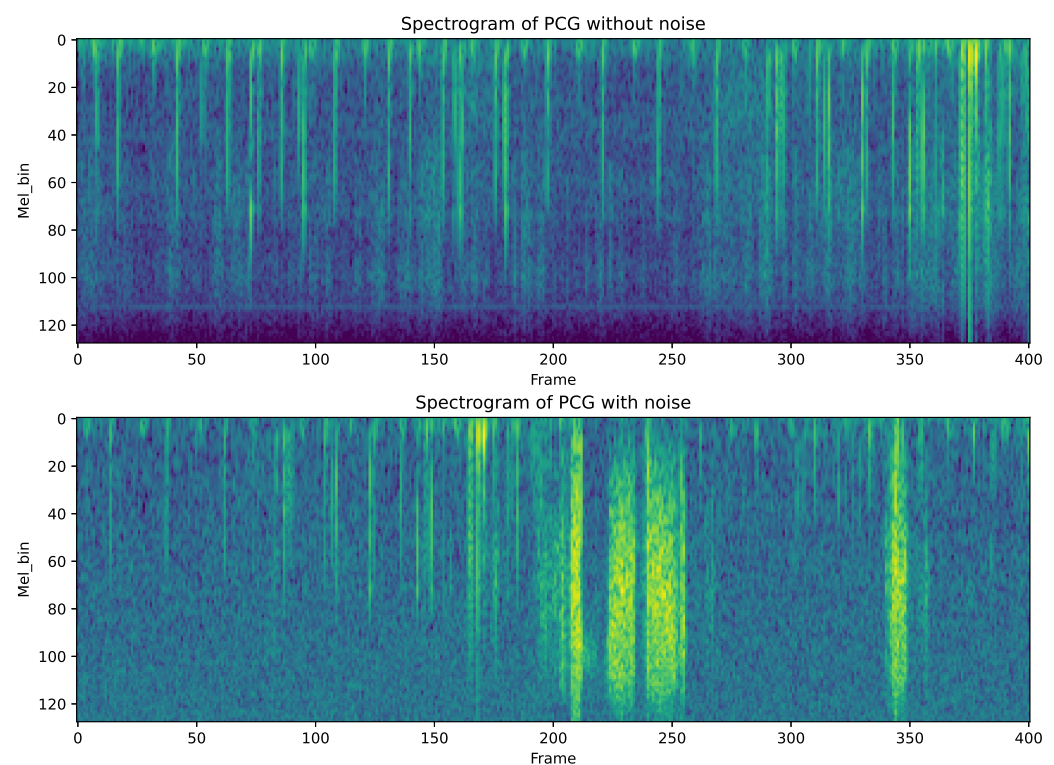


Figure 3. (Top): Mel-spectrogram of original PCG, (Bottom): Mel-spectrogram of PCG with Noise.

To enhance the robustness of our model against overfitting and to simulate real-world noisy environments, we implemented two primary data augmentation techniques on the Mel-spectrograms. Initially, Gaussian noise with a zero mean and a standard deviation of 15 dB was added to each recording during training, as shown in Figure 4, with a probability of 0.5 for each signal. Additionally, we applied masking in both frequency and time domains to the spectrograms, as shown in Figure 3. This involved selectively blocking out multiple blocks of frequency channels and time steps. These augmentation strategies were

designed to introduce variability and improve the generalizability of our model across diverse scenarios.

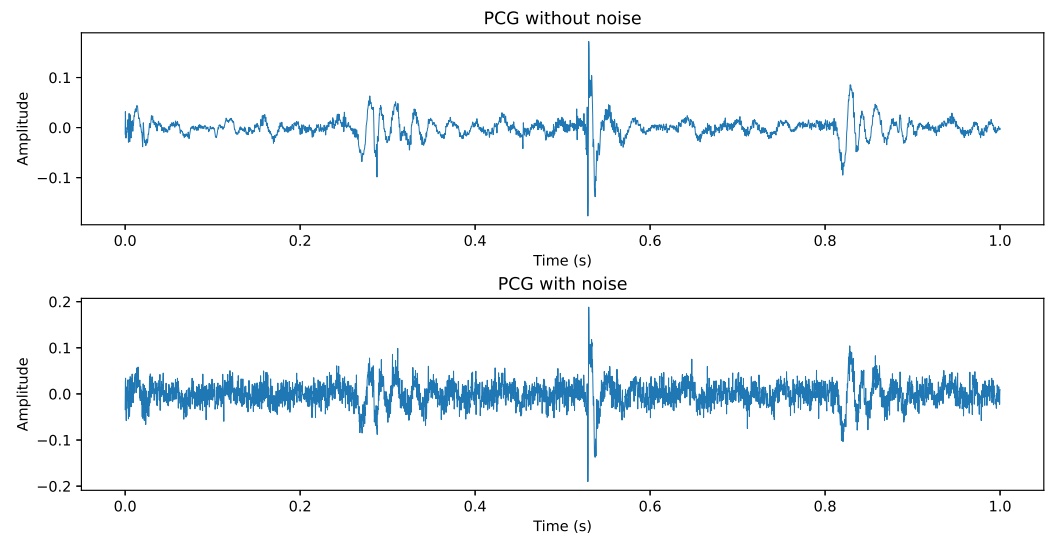


Figure 4. Top: Original PCG recording; Bottom: PCG recording after adding Gaussian noise.

4.2. Results

Due to significant class imbalance in the dataset, we used the murmur score from the PhysioNet Challenge 2022 [9] and AUC indicators to provide an unbiased evaluation of our model's performance. The murmur score is a weighted accuracy and can be defined by the confusion matrix shown in Table 2 and Formulation (4).

$$s_m = \frac{5m_{PP} + 3m_{UU} + m_{AA}}{5(m_{PP} + m_{UP} + m_{AP}) + 3(m_{PU} + m_{UU} + m_{AU}) + (m_{PA} + m_{UA} + m_{AA})} \quad (4)$$

To demonstrate the robustness and practicality of our methodology, we refer to Table 3, which presents both the average values and their corresponding standard deviations. Initially, the 2022 dataset was segmented into five folds, ensuring that all recordings from the same subject were confined to a single fold. This subject-wise cross-validation approach guaranteed that no recordings from the same patient appeared simultaneously in both the training and validation/test sets, thereby mitigating the risk of data leakage and providing a more reliable evaluation of the model performance. Using one fold for labeled training and another for validation, this approach achieved a five-fold cross-validation murmur score of 0.7020 and an AUC of 0.7951. Subsequently, applying our mean teacher model, we designated one fold labeled as training data while gradually increasing the unlabeled data from one fold to three folds, with the remaining fold reserved for validation. This semi-supervised strategy improved the model's performance, achieving a murmur score of 0.7540 and an AUC of 0.8562. As expected, we observed that as the amount of unlabeled data increased, the model's performance improved accordingly. The *t*-test show a *t*-value of -18.88 leading to a *p*-value of 1.36×10^{-4} on Murmur Score, and a *t*-value of -99.55 leading to a *p*-value of 7.79×10^{-8} on AUC. These results indicate that incorporating an additional 60% of unlabeled data significantly improved the model's performance over the baseline.

Table 2. Confusion matrix for murmur detection.

		Murmur Expert		
		Present	Unknown	Absent
Murmur Classifier	Present	m_{PP}	m_{PU}	m_{PA}
	Unknown	m_{UP}	m_{UU}	m_{UA}
	Absent	m_{AP}	m_{AU}	m_{AA}

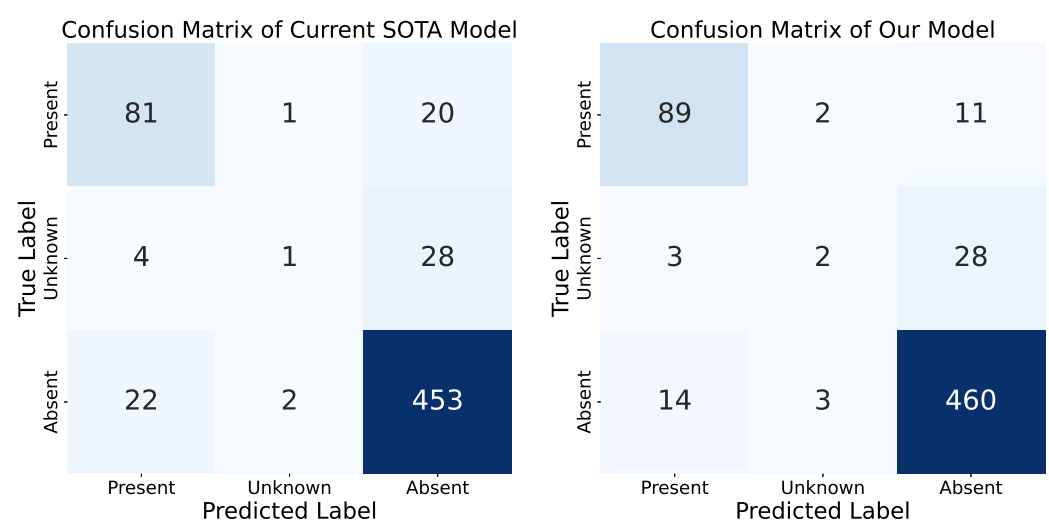
Table 3. Comparative analysis of various training modes across different data splits.

Model	Data Split			Result Measurements	
	Labeled Data	Unlabeled Data	Validation Data	Murmur Score	AUC
Full Supervise	20% v2022	/	20% v2022	0.7020 ± 0.0026	0.7951 ± 0.0008
Semi Supervise	20% v2022	20% v2022	20% v2022	0.7298 ± 0.0035	0.8217 ± 0.0009
Semi Supervise	20% v2022	40% v2022	20% v2022	0.7418 ± 0.0021	0.8398 ± 0.0013
Semi Supervise	20% v2022	60% v2022	20% v2022	0.7540 ± 0.0040	0.8562 ± 0.0007
Full Supervise (Current SOTA)	80% v2022	/	20% v2022	0.8074 ± 0.0010	0.8936 ± 0.0011
Semi Supervise (Ours)	80% v2022	100% v2016-e	20% v2022	0.8180 ± 0.0006	0.9004 ± 0.0019

Furthermore, we conducted an experiment using a five-fold cross-validation approach, where four folds were used as labeled data and the fifth fold served as the validation set. This setup resulted in a murmur score of 0.8074 and an AUC of 0.8936, highlighting the effectiveness and robustness of the method, and establishing it as the current SOTA approach in this domain.

To further enhance our semi-supervised learning strategy, we extended our approach by incorporating both the Physionet 2016 and 2022 challenge datasets. Specifically, we utilized 80% of the 2022 dataset as labeled data while using the entire 2016-e dataset as unlabeled data. The remaining 20% of the 2022 dataset was reserved for validation. This comprehensive configuration led to an improved murmur score of 0.8180 and an AUC of 0.9004. As shown in Figure 5, our model performs well in detecting murmurs in both the “absent” and “present” classes. However, it struggles significantly with the “unknown” class, which is the minority class in our dataset. This poor performance can also be attributed to the inherent clinical difficulty of the “unknown” class, even for clinical experts. Compared to the current SOTA model, our model significantly outperformed in both the present and absent class classifications.

These outcomes clearly demonstrate that our mean teacher method not only delivers more robust performance but also surpasses the results obtained by solely utilizing the 2022 challenge data. This underscores the effectiveness of incorporating unlabeled data to achieve a superior overall performance, highlighting the potential of our approach in advancing murmur detection and enhancing clinical decision-making in cardiovascular health. Consequently, our method establishes a new SOTA model for PCG murmur detection.

**Figure 5.** Comparison of the confusion matrix between the current SOTA model and our model.

5. Discussion

While our research demonstrates significant advancements in PCG murmur detection using a semi-supervised mean teacher method, certain limitations persist. Typically, semi-supervised learning frameworks are most effective when the volume of unlabeled data substantially exceeds that of the labeled data. In our study, the volume of unlabeled data was comparable to that of the labeled data, potentially capping the performance improvements that could be achieved. This suggests that acquiring a larger pool of unlabeled data could further enhance the model performance by better leveraging the strengths of semi-supervised learning. In our preprocessing strategy, recordings longer than 10 s were randomly cropped. Incorporating all available 10-second segments and applying additional data augmentation techniques could help to expand the dataset and improve performance. Future efforts could also focus on expanding the dataset, particularly with high-quality unlabeled PCG recordings, to fully exploit the potential of our semi-supervised approach and push the boundaries of the current SOTA method.

Additionally, the Physionet Challenge 2022 and 2016 datasets may have certain flaws that need to be addressed. Although expert labels are always considered the ground truths in our training and validation processes, emerging research has indicated that relabeling can significantly improve model performance [36]. This insight suggests that expert labels, while highly reliable, may still contain some inaccuracies that can impact the model's efficacy. A potential solution to this issue is to pinpoint ambiguous recordings within the dataset using the model's initial outputs. Once these ambiguous cases are identified, they can be relabeled with the assistance of expert knowledge. By double-checking and refining the labels, this approach can help to eliminate any discrepancies and improve the overall quality of the dataset. Besides, previous studies on the Physionet Challenge 2022 primarily concentrated on the performance of the murmur score [4–6,26]. While we have incorporated AUC performance in our experimental evaluation, other metrics such as sensitivity and specificity could also be considered for a more comprehensive assessment. Moreover, using additional high-quality PCG datasets, even simultaneous PCG and ECG (Electrocardiogram) dataset [37] for multi-signal source training may further elaborate our model. Finally, traditional murmur detection systems focus primarily on identifying the presence or absence of murmurs without differentiating their underlying cause or severity. However, heart murmurs, while indicative of abnormal heart activity, can stem from a variety of causes. Some of them are benign and others associated with serious valvular heart diseases such as aortic stenosis or mitral regurgitation. It is crucial clinically to distinguish between murmurs that are significant and those that are less concerning. As such, enhancing murmur detection algorithms to automatically classify murmurs based on their clinical relevance would also mark a significant step forward.

Also, the model's application to pediatric heart murmur detection and in LMICs presents certain limitations. As noted by [38], detecting heart murmurs in neonates and infants poses unique challenges due to the lower amplitude and higher variability of their heart sounds. Even though the PhysioNet2022 data focus on patients under 21 years old [9,39], for younger patients and children, the change in PCG signals would pose a challenge on our model. In addition, noisy environments are common in low-resource settings, where medical equipment may not meet optimal standards, can further degrade the model performance. The significant GPU computational requirements of the current model may also be difficult to meet in such environments. To address these challenges, future improvements should focus on creating pediatric-specific heart murmur datasets, enhancing the model's robustness to noise, and reducing its hardware and computational demands.

6. Conclusions

Our study showcases the mean teacher model in PCG signal murmur detection, achieving SOTA performance. By integrating both labeled and unlabeled data from the Physionet 2022 and 2016 challenges, we significantly enhanced detection accuracy, culminating in an impressive murmur score of 0.8180 and an AUC of 0.9004. These results not only affirm

our model's superiority over existing methods, but also demonstrate its effectiveness in managing the sparse datasets typical in medical applications. The study underscores the critical role of extending the dataset with additional unlabeled data, which is pivotal in advancing semi-supervised learning to achieve better performance and robustness in clinical diagnostics. Furthermore, our method provides an accessible, highly accurate, and cost-effective approach for early-stage CVD diagnosis, especially in LMICs, aiming to improve clinical equity worldwide.

Author Contributions: Conceptualization, Y.L., Z.F. and K.D.; methodology, Y.L. and Z.F.; software, Y.L., Z.F. and Y.D.; investigation, Y.L., Z.F., Y.D. and X.C.; writing—original draft, Y.L. and Z.F.; writing—review and editing, Y.L. and Z.F.; visualization, Y.L., Z.F., Y.D. and X.C.; supervision, K.D.; project administration, K.D.; funding acquisition, K.D. All authors have read and agreed to the published version of the manuscript.

Funding: Research reported in this publication was supported by the National Institutes of Health (award numbers R37CA229417).

Institutional Review Board Statement: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Informed Consent Statement: Informed consent for participation was not required, as the dataset used in this research is publicly available.

Data Availability Statement: The original data used in this study are openly available at <https://physionet.org/content/challenge-2016/1.0.0/> and <https://physionet.org/content/challenge-2022/1.0.0/>. The models presented in this study are available upon request from the corresponding author.

Acknowledgments: We acknowledge the invaluable guidance from Vishal Patel throughout the research.

Conflicts of Interest: The authors declare no conflict of interest.

References

- World Health Organization. *Cardiovascular Diseases (CVDs)*; World Health Organization: Geneva, Switzerland, 2017.
- Dornbush, S.; Turnquest, A.E. Physiology, heart sounds 2023. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, January 2024. [PubMed]
- Hoeting, N.M.; McCracken, C.E.; McConnell, M.; Sallee, D.; Iannucci, G.J.; Oster, M.E. Systolic ejection click versus split first heart sound: Are our ears deceiving us? *Congenit. Heart Dis.* **2017**, *12*, 417–420. [CrossRef] [PubMed]
- Lu, H.; Yip, J.B.; Steigleder, T.; Griebhammer, S.; Heckel, M.; Jami, N.V.S.J.; Eskofier, B.; Ostgathe, C.; Koelpin, A. A Lightweight Robust Approach for Automatic Heart Murmurs and Clinical Outcomes Classification from Phonocardiogram Recordings. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; Volume 498, pp. 1–4. [CrossRef]
- Xu, Y.; Bao, X.; Lam, H.K.; Kamavuako, E.N. Hierarchical Multi-Scale Convolutional Network for Murmurs Detection on PCG Signals. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; Volume 498, pp. 1–4. [CrossRef]
- McDonald, A.; Gales, M.J.; Agarwal, A. Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; Volume 498, pp. 1–4. [CrossRef]
- Cai, Q.; Pan, Y.; Ngo, C.W.; Tian, X.; Duan, L.; Yao, T. Exploring object relation in mean teacher for cross-domain detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11457–11466.
- Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1195–1204.
- Reyna, M.A.; Kiarashi, Y.; Elola, A.; Oliveira, J.; Renna, F.; Gu, A.; Perez Alday, E.A.; Sadr, N.; Sharma, A.; Kpodonu, J.; et al. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. *PLoS Digit. Health* **2023**, *2*, e0000324. [CrossRef] [PubMed]
- Clifford, G.D.; Liu, C.; Moody, B.; Springer, D.; Silva, I.; Li, Q.; Mark, R.G. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In Proceedings of the 2016 Computing in cardiology conference (CinC), Vancouver, BC, Canada, 11–14 September 2016; pp. 609–612.
- Dissanayake, T.; Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. Generalized generative deep learning models for biosignal synthesis and modality transfer. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 968–979. [CrossRef] [PubMed]
- Gerbarg, D.S.; Taranta, A.; Spagnuolo, M.; Hofler, J.J. Computer analysis of phonocardiograms. *Prog. Cardiovasc. Dis.* **1963**, *5*, 393–405. [CrossRef]

13. El-Segaier, M.; Lilja, O.; Lukkarinen, S.; Sörnmo, L.; Sepponen, R.; Pesonen, E. Computer-based detection and analysis of heart sound and murmur. *Ann. Biomed. Eng.* **2005**, *33*, 937–942. [[CrossRef](#)]
14. Delgado-Trejos, E.; Quiceno-Manrique, A.; Godino-Llorente, J.; Blanco-Velasco, M.; Castellanos-Dominguez, G. Digital auscultation analysis for heart murmur detection. *Ann. Biomed. Eng.* **2009**, *37*, 337–353. [[CrossRef](#)]
15. Ahlström, C. Processing of the Phonocardiographic Signal: Methods for the Intelligent Stethoscope. Ph.D. Thesis, Institutionen för Medicinsk Teknik, Linköping, Sweden, 2006.
16. Kumar, D.; Carvalho, P.; Antunes, M.; Henriques, J.; Melo, A.S.E.; Schmidt, R.; Habetha, J. Third Heart Sound Detection Using Wavelet Transform-Simplicity Filter. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; pp. 1277–1281. [[CrossRef](#)]
17. Eslamzadeh, G.; Barati, R. Heart murmur detection based on wavelet transformation and a synergy between artificial neural network and modified neighbor annealing methods. *Artif. Intell. Med.* **2017**, *78*, 23–40. [[CrossRef](#)]
18. Varghees, V.N.; Ramachandran, K. Effective heart sound segmentation and murmur classification using empirical wavelet transform and instantaneous phase for electronic stethoscope. *IEEE Sens. J.* **2017**, *17*, 3861–3872. [[CrossRef](#)]
19. Potes, C.; Parvaneh, S.; Rahman, A.; Conroy, B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 11–14 September 2016; pp. 621–624.
20. Chorba, J.S.; Shapiro, A.M.; Le, L.; Maidens, J.; Prince, J.; Pham, S.; Kanzawa, M.M.; Barbosa, D.N.; Currie, C.; Brooks, C.; et al. Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. *J. Am. Heart Assoc.* **2021**, *10*, e019905. [[CrossRef](#)]
21. Abduh, Z.; Nehary, E.A.; Wahed, M.A.; Kadah, Y.M. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers. *Biomed. Signal Process. Control* **2020**, *57*, 101788. [[CrossRef](#)]
22. 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. *Comput. Biol. Med.* **2020**, *120*, 103733. [[CrossRef](#)] [[PubMed](#)]
23. Megalmani, D.R.; Shailesh, B.; Rao, A.; Jeevannavar, S.S.; Ghosh, P.K. Unsegmented heart sound classification using hybrid CNN-LSTM neural networks. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 31Oct–4 November 2021; pp. 713–717.
24. Bao, X.; Xu, Y.; Kamavuako, E.N. The effect of signal duration on the classification of heart sounds: A deep learning approach. *Sensors* **2022**, *22*, 2261. [[CrossRef](#)] [[PubMed](#)]
25. Almanifi, O.R.A.; Ab Nasir, A.F.; Razman, M.A.M.; Musa, R.M.; Majeed, A.P.A. Heartbeat murmurs detection in phonocardiogram recordings via transfer learning. *Alex. Eng. J.* **2022**, *61*, 10995–11002. [[CrossRef](#)]
26. Lee, J.; Kang, T.; Kim, N.; Han, S.; Won, H.; Gong, W.; Kwak, I.Y. Deep learning based heart murmur detection using frequency-time domain features of heartbeat sounds. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; Volume 498, pp. 1–4.
27. Walker, B.; Krones, F.; Kiskin, I.; Parsons, G.; Lyons, T.; Mahdi, A. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; Volume 498, pp. 1–4.
28. Panah, D.S.; Hines, A.; McKeever, S. Exploring wav2vec 2.0 model for heart murmur detection. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; pp. 1010–1014.
29. Kamson, A.P.; Lewis, M.C.; Sunil, B.V.; Jeevannavar, S.S.; Sawant, A.; Ghosh, P.K. E2E Multi-Scale CNN with LSTM for Murmur Detection in PCG or Noise Identification. In Proceedings of the 2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE), Dubai, United Arab Emirates, 30–31 December 2023; pp. 1–6.
30. Bondareva, E.; Rizos, G.; Han, J.; Mascolo, C. Embracing the Imaginary: Deep Complex-valued Networks for Heart Murmur Detection. In Proceedings of the 2023 Computing in Cardiology (CinC), Atlanta, GA, USA, 1–4 October 2023; Volume 50, pp. 1–4.
31. Rohr, M.; Müller, B.; Dill, S.; Güney, G.; Hoog Antink, C. Multiple instance learning framework can facilitate explainability in murmur detection. *PLoS Digit. Health* **2024**, *3*, e0000461. [[CrossRef](#)] [[PubMed](#)]
32. Zhou, G.; Chen, Y.; Chien, C. On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks. *BMC Med Inform. Decis. Mak.* **2022**, *22*, 226. [[CrossRef](#)]
33. Bachman, P.; Alsharif, O.; Precup, D. Learning with pseudo-ensembles. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3365–3373.
34. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
35. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
36. Patwa, A.; Rahman, M.M.U.; Al-Naffouri, T.Y. Heart murmur and abnormal pcg detection via wavelet scattering transform & a 1d-cnn. *arXiv* **2023**, arXiv:2303.11423.
37. Kazemnejad, A.; Gordany, P.; Sameni, R. EPHNOGRAM: A simultaneous electrocardiogram and phonocardiogram database. *PhysioNet* **2021**. [[CrossRef](#)]

38. Nojonen, A.L.; Lukkarinen, S.; Angerla, A.; Sepponen, R. Phono-spectrographic analysis of heart murmur in children. *BMC Pediatr.* **2007**, *7*, 1–10. [[CrossRef](#)] [[PubMed](#)]
39. Oliveira, J.; Renna, F.; Costa, P.D.; Nogueira, M.; Oliveira, C.; Ferreira, C.; Jorge, A.; Mattos, S.; Hatem, T.; Tavares, T.; et al. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 2524–2535. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.