Listen to your heart: A self-supervised approach for detecting murmur in heart-beat sounds for the Physionet 2022 challenge

Aristotelis Ballas[1], Vasileios Papapanagiotou[2], Anastasios Delopoulos[2], Christos Diou[1]

[1] Harokopio University of Athens, Athens, Greece
[2] Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

Heart murmurs are abnormal sounds present in heartbeats, caused by turbulent blood flow through the heart. The PhysioNet 2022 challenge targets automatic detection of murmur from audio recordings of the heart and automatic detection of normal vs. abnormal clinical outcome. The recordings are captured from multiple locations around the heart.

Our participation investigates the effectiveness of self-supervised learning for murmur detection. We evaluate the use of a backbone CNN, whose layers are trained in a self-supervised way with data from both this year's and the 2016 challenge. We use two different augmentations on each training sample, and normalized temperature-scaled cross-entropy loss. We experiment with different augmentations to learn effective phonocardiogram representations. To build the final detectors we train two classification heads, one for each challenge task. We present evaluation results for all combinations of the available augmentations, and for our multiple-augmentation approach.

1. Introduction

Heart murmurs can be a sign of a number of serious heart conditions associated with increased mortality [1]. The development of accurate, non-invasive and easily accessible methods for heart murmur screening can help with prevention and/or timely treatment, especially for populations with reduced access to healthcare resources. The 2022 PhysioNet Challenge [2, 3] attempts to incentivize researchers towards developing methods for the automatic detection of abnormal heart waves or heart murmurs, by analyzing phonocardiogram (PCG) signals. The analysis of such auscultation signals can provide crucial information on the heart's structure and potentially assist clinicians in early screenings or assessments. We propose tackling the challenge through Self-Supervised Learning (SSL) with Deep Neural Networks (DNN) for the classification of PCG signals.

Although DNNs have achieved remarkable success in several benchmarks and experiments, they have not yet been able to establish themselves in clinical practice (including screening procedures). There are several reasons that may contribute to this. First of all, a false prediction may lead to serious consequences for the patient's health or may incur unnecessary healthcare costs. Secondly, an already trained model may not maintain its performance on newly collected data, which was collected using different procedures, devices or by distinct populations [4]. An additional constraint is the fact that hospital and medical data may be weakly labeled, unlabeled and may be affected by exogenous noise.

All the above add up to a need for generalizable, adaptable and non-conventional model training methods. To this end, Self-Supervised Learning (SSL) techniques have been proposed for biosignal classification and have achieved promising results [5]. In SSL, the main idea is to learn representations by transforming or augmenting the input data and by training a model to distinguish between representations that originate from the same signal against those that come from different ones. The model, therefore, learns a representation of the input data which remains invariant to these transformations and hopefully captures its informative features. In addition to medical image analysis [6], SSL methods have also been implemented for EEG [7, 8] and ECG [9, 10] signals. In this work, we explore the usage of contrastive SSL [11] for the classification of PCG signals. Specifically, we (i) propose employing SSL for the extraction of meaningful PCG signal representation, (ii) evaluate our method's effectiveness with minimal fine-tuning on downstream tasks and (iii) explore promising combinations of augmentations for the PCG signal domain.

2. Materials and methods

The goal of the challenge [2, 3] is to classify a patient on two classification tasks: murmur detection and clinical outcome. For each patient there are one or more audio

recordings available, taken from 5 different locations (one of the four valves or other).

Our approach is split in two steps, and is similar for each classification task. First, we train an audio-based classifier on windows of fixed length by directly propagating the patient label to the windows (the only exception is for the murmur class "present" where only windows from recording locations marked as "murmur locations" are labeled as "present"; windows from the rest recording locations are labeled as "absent"). Then, we predict a single label for the patient by first aggregating the window-level labels to recording-level labels and then the recording-level labels to a single patient-level label.

Our audio-based classifier is a convolutional neural network, based on the architecture of [12]. However, given the limited size of the dataset, we train the convolutional layers in a self-supervised way, based on the approach of [13]. Thus, we also use the dataset released for the 2016 challenge of PhysioNet [14] to extend our training dataset. It should be noted that the 2016 dataset does not contain labels that are relevant to the 2022 challenge.

The 2022 dataset contains audio at 4 kHz while the 2016 dataset at 2 kHz. Therefore, we choose to resample the 2022 dataset at 2 kHz. We use windows of 5 sec length with an overlap of 2.5 sec. However, the first and last 2 seconds of each recording are discarded as they might contain transient noise artifacts.

2.1. Self-supervised training

Each audio window x is transformed into two versions, namely x_1 and x_2, by applying two different sets of augmentations. Specifically, we examine the following augmentations: cut-off filters (high-pass or low-pass, at 250, 500, and 750 Hz), rewinding the signal, inverting (multiply by -1), random scaling (in the range of 0.5 to 2.0), adding zero-mean uniform noise and upsampling by a factor of 2 (and cropping symmetrically to maintain constant window size in samples).

To train the 5 convolutional layers of the 5-sec input window architecture of [12] we project the output of the last convolutional layer on an 128-D space (via a fully connected layer) and then minimize the normalized temperature-scaled cross-entropy loss [15]. In detail, for the cosine similarity with temperature [16], we use a temperature value of 0.1 based on some initial experiments on the training set and on the findings of [13].

Batch size is set to 256 (leading to 512 after the two augmentations) and training is allowed for at most 50 epochs (we use early stopping with a patience of 5 epochs on a validation subset, more on Section 3.1). To train such a large batch we employ the LARS optimizer [17] with a linear warm-up of 5 epochs to a learning rate of 0.1 and then a cosine decay (with an alpha of 1%).

Once training concludes, the projection head (i.e. the FC layer with the 128 neurons) is discarded, and the weights of the convolutional layers are frozen.

2.2. Murmur classification

Given a trained set of convolutional layers, we create a murmur classifier by appending a classification head of 3 FC layers (including dropout of 0.5 probability) as in [13]. The last layer has 3 neurons, one per murmur class ("present", "unknown", and "absent") and uses softmax activation. We train only the FC layers using categorical cross-entropy loss and the Adam optimizer, with a learning rate of 10^{-4} (selected after some initial experimentations with the training set). Batch size is set to 64 and we train for at most 200 epochs using early stopping with 20 epochs patience on a validation subset (more on Section 3.1).

Training yields a murmur classifier for audio windows. Given an audio recording, we extract windows of 5-second length and 2.5-second overlap (identically to the training process) and apply the trained murmur classifier on each window independently. We thus infer the probability of each murmur class for each window. These probabilities are aggregated for the entire recording by averaging across the windows (we have experimented with other aggregations but this method evidently seems to be the most effective). Ultimately, we characterize the entire recording with a single label, based on the most probable class.

To assign a "present" label to a patient, we require that at least one recording has been labeled "present". If none of the recordings have been labeled as "present" and at least one recording has been labeled as "unknown" the patient inherits the "unknown" label. Finally, the "absent" label is assigned to the patient only in the case that all recordings have been labeled accordingly.

2.3. Outcome classification

Training and inference for clinical outcome are very similar to murmur training and inference respectively. The only difference is that the classification problem is now binary ("abnormal" vs. "normal" clinical outcome). Window-level aggregations are performed again using averaging, and we require that at least one recording is classified as "abnormal" in order to label the patient's clinical outcome as abnormal. If no such recording exists, we label the patient as "normal".

3. Experiments and results

3.1. Experimental setup

The dataset provided by the challenge contains data from 942 patients. We perform a stratified random split

and keep 100 patients as our test set, denoted S_{test} (different from the hidden test set of the challenge). The remaining 842 patients are again split (using a stratified random split) into a training set S_{train} and a validation set S_{val} (80%-20%). For SSL training, we also use the publically available 2016 challenge dataset [14, 18], which contains $3,541$ audio recordings, denoted S_{16}. We would like to note that S_{16} does not contain labels relevant to this year's challenge. Thus, SSL training is done on $S_{16} \cup S_{\text{train}}$, downstream training is done on S_{train}, and validation is always S_{val}. All results are reported on S_{test}.

3.2. Evaluation of augmentations

To adequately explore the most promising augmentations, we first evaluate each one of them separately (i.e. $x_1 = f_i(x)$ and $x_2 = x$ where f_i are the different augmentations we examine) and subsequently in pairs (i.e. $x_1 = f_i(x)$ and $x_2 = f_j(x)$). Figure 1 summarizes the results for classification accuracy both for murmur and outcome labels.

For murmur classification and a single augmentation, the first column of Figure 1 shows that a low-pass filter at 250 Hz, rewind, and uniform noise (of 0.02 range) yield the best results (accuracy over 0.8). The highest accuracy for pairs is 0.85 and is obtained by two combinations: a high-pass filter at 500 Hz with a high-pass filter at 250 Hz, and random scaling with itself. Both of these pairs have the same or similar augmentations, however, and the resulting representations may not be informative. The following combinations of diverse augmentations achieve 0.84 accuracy: invert with rewind, low-pass filter at 500 Hz with uniform noise (range 0.002), random scale with uniform noise (range 0.002).

Accuracy for clinical outcome is generally lower. This is most probably due to the fact that, in clinical practice, decision for clinical outcome does not rely solely on listening to the audio recordings but also considers additional information such as the patient/examinee's medical history (which were not taken into account in our case). The best results are obtained by combining uniform noise (range 0.01) with either a high-pass filter at 750 Hz, yielding 0.65 accuracy, or with upsampling by 2, yielding 0.64 accuracy.

3.3. Challenge results

Based on these findings, we explore adding a second augmentation to one or both of the window copies. The final approach (that was submitted to the challenge) performs the following augemntations: the first copy of the window is high-pass filtered at 250 Hz (f_1), followed by rewind with a 0.5 probability (f_2), followed by inverting with a 0.5 probability (f_3), while the second copy is polluted with uniform noise with a range of 0.02 (g_1) followed by upsampling with a 0.5 probability (g_2).

We present the results for this setup on the first row of Table 1, while also reporting the following additional metrics: F1-score, accuracy, and weighted accuracy. It should be noted that accuracy is weighed based on the challenge guidelines (for murmur: "present" = 5, "unknown" = 3, and "absent" = 1, for outcome: "abnormal" = 5 and "normal" = 1). In the following two lines we present similar results where we omit one augmentation from each channel (rewind and inversion are removed together). Finally, the last line reports baseline results where no SSL is used at all, and the network is trained from top to bottom (no frozen weights) and individually for the murmur and outcome tasks solely on the 2022 dataset.

The challenge is ranked based on murmur weighted accuracy and clinical outcome cost. Our models are able to surpass the fully supervised baseline and achieve a weighted accuracy of 0.668 for murmur classification and a cost of $10,541$ for clinical outcome prediction. On the challenge's hidden test set, we scored 0.671 for murmur weighted accuracy and $10,821$ for outcome cost.

4. Conclusions

This work approaches the problem of PCG classification through SSL and evaluates its performance on the PhysioNet 2022 challenge. Results indicate that a backbone network trained via contrastive SSL improves performance compared to end-to-end supervised network training. By exploring and combining several augmentation techniques we are able to gain some intuition into which data transformations assist model training in this particular signal domain. Following this research, we aim to explore additional transformations and to introduce attention mechanisms into the model.

Acknowledgments

References

[1] GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet November 2018;392(10159):1736–1788. ISSN 1474-547X.

[2] Reyna M, et al. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. medRxiv 2022;.
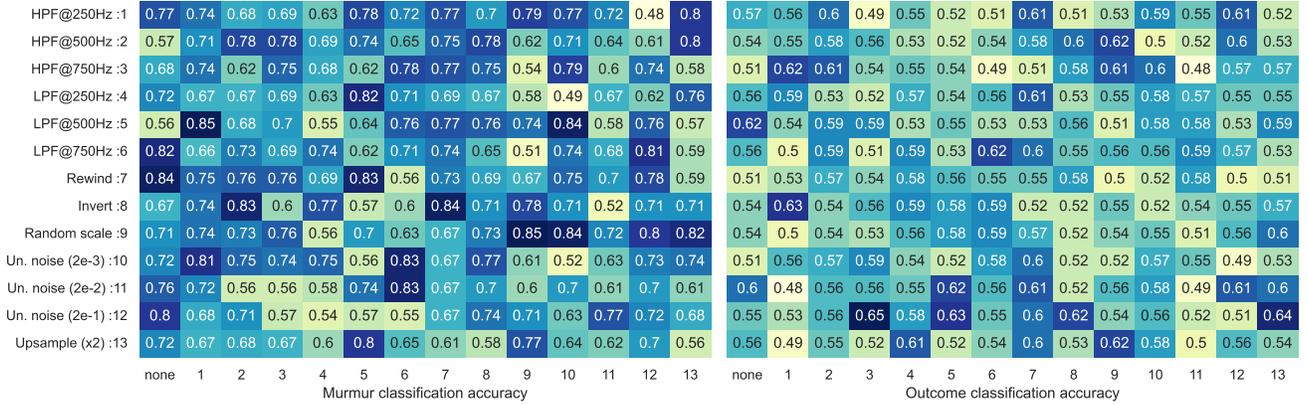
Figure 1: Augmentations comparison based on accuracy for murmur (left) and outcome (right) classification tasks.

Murmur classification accuracy

Augmentation	none	1	2	3	4	5	6	7	8	9	10	11	12	13
HPF@250Hz :1	0.77	0.74	0.68	0.69	0.63	0.78	0.72	0.77	0.7	0.79	0.77	0.72	0.48	0.8
HPF@500Hz :2	0.57	0.71	0.78	0.78	0.69	0.74	0.65	0.75	0.78	0.62	0.71	0.64	0.61	0.8
HPF@750Hz :3	0.68	0.74	0.62	0.75	0.68	0.62	0.78	0.77	0.75	0.54	0.79	0.6	0.74	0.58
LPF@250Hz :4	0.72	0.67	0.67	0.69	0.63	0.82	0.71	0.69	0.67	0.58	0.49	0.67	0.62	0.76
LPF@500Hz :5	0.56	0.85	0.68	0.7	0.55	0.64	0.76	0.77	0.76	0.74	0.84	0.58	0.76	0.57
LPF@750Hz :6	0.82	0.66	0.73	0.69	0.74	0.62	0.71	0.74	0.65	0.51	0.74	0.68	0.81	0.59
Rewind :7	0.84	0.75	0.76	0.76	0.69	0.83	0.56	0.73	0.69	0.67	0.75	0.7	0.78	0.59
Invert :8	0.67	0.74	0.83	0.6	0.77	0.57	0.6	0.84	0.71	0.78	0.71	0.52	0.71	0.71
Random scale :9	0.71	0.74	0.73	0.76	0.56	0.7	0.63	0.67	0.73	0.85	0.84	0.72	0.8	0.82
Un. noise (2e-3) :10	0.72	0.81	0.75	0.74	0.75	0.56	0.83	0.67	0.77	0.61	0.52	0.63	0.73	0.74
Un. noise (2e-2) :11	0.76	0.72	0.56	0.56	0.58	0.74	0.83	0.67	0.7	0.6	0.7	0.61	0.7	0.61
Un. noise (2e-1) :12	0.8	0.68	0.71	0.57	0.54	0.57	0.55	0.67	0.74	0.71	0.63	0.77	0.72	0.68
Upsample (x2) :13	0.72	0.67	0.68	0.67	0.6	0.8	0.65	0.61	0.58	0.77	0.64	0.62	0.7	0.56

Outcome classification accuracy

Augmentation	none	1	2	3	4	5	6	7	8	9	10	11	12	13
HPF@250Hz :1	0.57	0.56	0.6	0.49	0.55	0.52	0.51	0.61	0.51	0.53	0.59	0.55	0.61	0.52
HPF@500Hz :2	0.54	0.55	0.58	0.56	0.53	0.52	0.54	0.58	0.6	0.62	0.5	0.52	0.6	0.53
HPF@750Hz :3	0.51	0.62	0.61	0.54	0.55	0.54	0.49	0.51	0.58	0.61	0.6	0.48	0.57	0.57
LPF@250Hz :4	0.56	0.59	0.53	0.52	0.57	0.54	0.56	0.61	0.53	0.55	0.58	0.57	0.55	0.55
LPF@500Hz :5	0.62	0.54	0.59	0.59	0.53	0.55	0.53	0.53	0.56	0.51	0.58	0.58	0.53	0.59
LPF@750Hz :6	0.56	0.5	0.59	0.51	0.59	0.53	0.62	0.6	0.55	0.56	0.56	0.59	0.57	0.53
Rewind :7	0.51	0.53	0.57	0.54	0.58	0.56	0.55	0.55	0.58	0.5	0.52	0.58	0.5	0.51
Invert :8	0.54	0.63	0.54	0.56	0.59	0.58	0.59	0.52	0.52	0.55	0.52	0.54	0.55	0.57
Random scale :9	0.54	0.5	0.54	0.53	0.56	0.58	0.59	0.57	0.52	0.54	0.55	0.51	0.56	0.6
Un. noise (2e-3) :10	0.51	0.56	0.57	0.59	0.54	0.52	0.58	0.6	0.52	0.52	0.57	0.55	0.49	0.53
Un. noise (2e-2) :11	0.6	0.48	0.56	0.56	0.55	0.62	0.56	0.61	0.52	0.56	0.58	0.49	0.61	0.6
Un. noise (2e-1) :12	0.55	0.53	0.56	0.65	0.58	0.63	0.55	0.6	0.62	0.54	0.56	0.52	0.51	0.64
Upsample (x2) :13	0.56	0.49	0.55	0.52	0.61	0.52	0.54	0.6	0.53	0.62	0.58	0.5	0.56	0.54

Table 1: Final results and baseline (fully supervised training on S_{train}).

	murmur			outcome			
	F1-score	acc.	w. acc.	F1-score	acc.	w. acc.	cost
$f_3(f_2(f_1))$ vs. $g_2(g_1)$	0.544	0.590	0.668	0.657	0.590	0.764	10,541
f_1 vs. $g_2(g_1)$	0.606	0.670	0.700	0.531	0.570	0.784	10,892
$f_3(f_2(f_1))$ vs. g_1	0.561	0.650	0.700	0.495	0.540	0.760	11,513
Baseline	0.524	0.600	0.558	0.515	0.550	0.750	11,276

[3] Oliveira J, et al. The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. IEEE Journal of Biomedical and Health Informatics 2022;26(6):2524–2535.

[4] Ballas A, Diou C. A domain generalization approach for out-of-distribution 12-lead ecg classification with convolutional neural networks. In 2022 IEEE Eight International Conference on Big Data Computing Service and Applications (BigDataService). 2022; ArXiv: 2208.09656.

[5] Spathis D, Perez-Pozuelo I, Marques-Fernandez L, Mascolo C. Breaking away from labels: The promise of self-supervised machine learning in intelligent health. Patterns 2022;3(2):100410. ISSN 2666-3899.

[6] Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis 2019;58:101539.

[7] Gramfort A, Banville H, Chehab O, Hyvärinen A, Engemann D. Learning with self-supervision on EEG data. In 2021 9th International Winter Conference on Brain-Computer Interface (BCI). 2021; .

[8] Krell MM, Kim SK. Rotational data augmentation for electroencephalographic data. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2017; 471–474.

[9] Sarkar P, Etemad A. Self-supervised ECG Representation Learning for Emotion Recognition. IEEE Trans Affective Comput 2021;ISSN 1949-3045, 2371-9850. ArXiv: 2002.03898.

[10] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. arXiv210312676 cs eess stat March 2021;ArXiv: 2103.12676.

[11] Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A Survey on Contrastive Self-Supervised Learning. Technologies March 2021;9(1):2. ISSN 2227-7080. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[12] Papapanagiotou V, Diou C, Delopoulos A. Chewing detection from an in-ear microphone using convolutional neural networks. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2017; 1258–1261.

[13] Papapanagiotou V, Diou C, Delopoulos A. Self-Supervised Feature Learning of 1D Convolutional Neural Networks with Contrastive Loss for Eating Detection Using an In-Ear Microphone. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2021; 7186–7189.

[14] Liu C, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement nov 2016;37(12):2181–2213.

[15] Oord Avd, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding. arxiv180703748 csLG 2018;.

[16] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. arXiv150302531 statML 2015;.

[17] You Y, Gitman I, Ginsburg B. Large Batch Training of Convolutional Networks. arXiv170803888 csCV 2017;.

[18] Goldberger A, Amaral L, L. G, Hausdorff J, Ivanov PC, Mark R, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals., 2000.