# Phonocardiogram Classification Using 1-Dimensional Inception Time Convolutional Neural Networks

Bjørn-Jostein Singstad[1], Antony M Gitau[2], Markus Kreutzer Johnsen[3], Johan Ravn[3], Lars Ailo Bongo[4], Henrik Schirmer[5]

[1] Simula Research Laboratory, Department of Computational Physiology, Oslo, Norway
[2] Kenyatta University, Nairobi, Kenya
[3] Medsensio AS, Oslo, Norway
[4] UiT, The Arctic University of Norway, Tromsø, Norway
[5] Akershus University Hospital, Lørenskog, Norway

## Abstract

*Murmurs are sounds caused by turbulent blood flow that are often the first sign of structural heart disease. These sounds are detected by auscultating the heart using a stethoscope, or more recently by a phonocardiogram (PCG). We aim to identify the presence, absence, or unclear cases of murmurs, as well as predict normal or abnormal clinical outcome from PCG recordings using machine learning.*

*We trained and tested two 1-dimensional convolutional neural networks (CNN) on a PCG data set from a pediatric population of* $1568$ *individuals. One model predicted murmurs, while the other model predicted clinical outcomes. Both models were trained to give recording-wise predictions, while the final predictions were given for every patient (patient wise predictions).*

*This paper describes our participation in the George B. Moody PhysioNet Challenge 2022 whose objective was to identify heart murmurs and clinical outcome from PCGs. Our team, Simulab, trained a clinical outcome classifier that achieved a challenge cost score of* $12419$ *(ranked 14th out of 39 teams) and the murmur classifier achieved a weighted accuracy of* $0.593$ *(ranked* $30th$ *out of* $40$ *teams) on the test set.*

## 1. Introduction

Cardiovascular diseases are one of the major causes of death, and represent $32\%$ of all global deaths [1]. Heart sounds provide an important source of information to a clinician in detecting abnormal murmurs which might be a sign of structural heart disease [2]. The most common and cost-effective tool for acquiring heart sounds is the stethoscope [3]. Despite the advancement of new cardiac monitoring methods, the stethoscope still remains an important tool for first-line cardiac screening, when performed correctly [4]. However, studies show that auscultation using a stethoscope is generally poorly performed both by medical students [5] and physicians [6], and many physicians cannot reliably distinguish abnormal from normal heart sounds, especially in children [7].

A phonocardiogram (PCG) is a digital representation of a heart sound and can be recorded by a phonocardiograph. A phonocardiograph is a stethoscope that transmits the sounds to a digital sampling device instead of transmitting them to the clinician's ears like a stethoscope. Figure 1 show an example of a PCG plotted in the time domain, and also the two most prominent peaks in a cardiac cycle which are called the first heart sound (S1) and the second heart sound (S2). S1 originates from the closing of mitral and tricuspid valves after blood flows from atria to ventricles, while S2 is caused by the closing of the aortic and pulmonary valve after blood is ejected from the two ventricles. Heart sounds are usually auscultated at the four different locations on the chest wall which corresponds to the aortic, pulmonary, tricuspid and mitral valves.

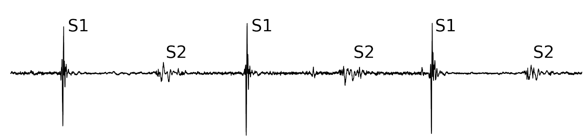


Figure 1. An example of a phonocardiogram showing three cardiac cycles

Back in 1987 Rangayyan, R. M., & Lehner, R. J. stated that: "The heart sound signal has much more information than can be assessed by the human ear or by visual inspection of the signal tracings on paper as currently practiced" [8] and since then, several attempts have been made to pro-

 ISSN: 2325-887X DOI: 10.22489/CinC.2022.108

cess [9], analyse [10] and classify PCG recordings using deep learning methods [11].

This paper describes our approach towards achieving the objective in George B. Moody PhysioNet Challenge 2022. We report our methodology and results and from developing two algorithms that identify murmurs as present, absent or unclear and classify the clinical outcome of a patient as normal or abnormal from PCG. Furthermore, we discuss the results and end our paper with a conclusion.

## 2. Method

We use a supervised machine learning approach, a convolutional neural network (CNN), to detect murmurs and classify the clinical outcome of a patient using a single PCG signal. We implemented the models in Python (3.8.9) using Tensorflow (2.8.2). The code is open sourced and published on GitHub [1].

### 2.1. Data

The data set used in this work consists of $5272$ PCGs from a pediatric population of $1568$ individuals [4, 12]. $3163$ PCGs from $942$ individuals were used for training. The remaining $2109$ PCGs from $149$ and $477$ patients, only available to the organizers of the challenge, were used for validation and testing respectively. Each patient could have one or more PCG recordings taken from a location close to the aortic valve, pulmonary valve, tricuspid valve, mitral valve or in some cases unknown.

Each patient was labeled with a clinical outcome (abnormal/normal) and murmur (present/unknown/absent), annotated by a clinical expert [3]. In cases of present murmur, the location of the recorded murmur was given in the training set.

### 2.2. Pre-processing

#### 2.2.1. Signal processing

The PCG signals in the training data were recorded with a sampling frequency of 4000Hz. We downsampled all signals to 100Hz. In addition, we zero padded all signals such that all signals were of a length equal to the longest signal in the training data, which was 6451 samples after down sampling to 100Hz. 6451 samples were also used as the length threshold for the validation and test data after down sampling to 100Hz. Signals with length $l < 6451$ were given a zero-padded tail of length $6451 - l$ and signals longer than $6451$ were truncated.

[1]This link will be valid after the challenge is finished: https://github.com/Bsingstad/Heart-murmur-detection-2022-private

#### 2.2.2. Label processing

The data set was relabeled from patient-wise labeling to recording-wise labeling. This was done by labeling all PCGs from a patient with the same clinical outcome as the original overall label. The recording-wise relabeling of murmurs is shown in Algorithm 1.

**Algorithm 1** : Patient to recording wise murmur labels

**Input:** $p$ = patient, $r$ = PCG record,
$t$ = total population, $l$ = label
**Output:** $r_l$ = recording wise labels
**for** $n$ in $t$ **do**
  **if** $p_{n_l} = Absent$ **then**
    **all** $r_l$ **in** $p_n = Absent$
  **else if** $p_{n_l} = Unknown$ **then**
    **all** $r_l$ **in** $p_n = Unknown$
  **else if** $p_{n_l} = Present$ **then**
    **for** $m$ **in** $p_{n_r}$ **do**
      $r_{l_m} = Present$
    **end for**
  **end if**
**end for**

### 2.3. Models

Two classification models were trained; one model to classify murmurs and the other to classify outcomes. Both models were 1 dimensional CNNs with an Inception Time architecture [13]. The murmur model was a multi-class classifier, set to classify whether a murmur was present/absent/unknown in the heart sound recording. The outcome model was a binary classifier used to classify whether the patient would have a normal or abnormal outcome. The murmur classifier was trained using weighted categorical cross entropy, while the clinical outcome classifier was trained using weighted binary cross entropy. The weights in both models were determined to be inversely proportional to the prevalence of the classes.

### 2.4. Post-processing

The recording-wise predictions from the model were finally converted back to patient-wise predictions. The murmur conversion is shown in Algorithm 2 and the clinical outcome conversion is shown in Algorithm 3.

### 2.5. Model selection (local development)

To estimate the performance of the models we did local training and validation on the training set, using using 5-fold cross-validation (CV), using 5-fold cross-validation (CV), before submitting the final model to the organizers

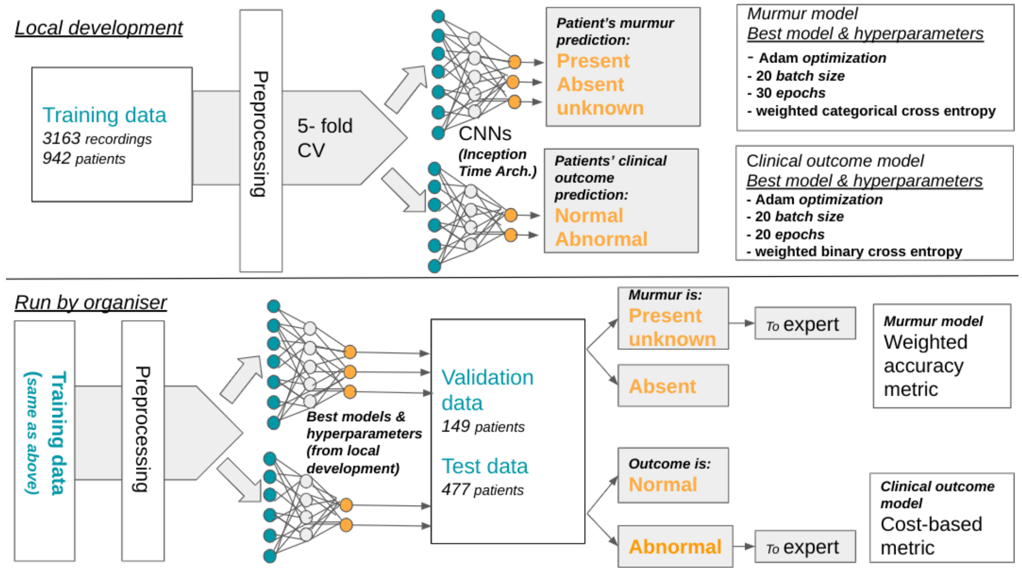


Figure 2. Detailed overview of the local development of the model and the submitted code run by the organizers.

of the challenge. The CV folds were stratified on patient level, and new models were trained and validated each successive round.

**Algorithm 2** : Murmur algorithm

**Input:** $p$ = patient, $r$ = PCG record,
$t$ = total population, $l$ = label
**Output:** $p_{n_l}$ = patient wise labels
**for** $n$ in $t$ **do**
  **if any** $r_l$ **in** $p_n$ = Absent **then**
    $p_{n_l}$ = Absent
  **else if any** $r_l$ **in** $p_n$ = Present **then**
    $p_{n_l}$ = Present
  **else if any** $r_l$ **in** $p_n$ = Unknown **then**
    $p_{n_l}$ = Unknown
  **end if**
**end for**

**Algorithm 3** : Outcome algorithm

**Input:** $p$ = patient, $r$ = PCG record,
$t$ = total population, $l$ = label
**Output:** $p_{n_l}$ = patient wise labels
**for** $n$ in $t$ **do**
  **if any** $r_l$ **in** $p_n$ = Abnormal **then**
    $p_{n_l}$ = Abnormal
  **else if any** $r_l$ **in** $p_n$ = Normal **then**
    $p_{n_l}$ = Normal
  **end if**
**end for**

## 2.6. Submitted model

The best models and hyper-parameters from local development were submitted to the organizers using a Docker image. The models were trained on the public training set and then applied to the hidden validation set and the best preforming model on the validation set were finally applied on the test set.

The murmur classifier was trained for 30 epochs while the clinical outcome classifier was trained for 20 epochs. Both models were trained using a batch size of 20, and an Adam optimizer starting at a learning rate = 0.001.

# 3. Results

Table 1 shows the CV results on the training data set as well as the results on the hidden validation and test set. The ranking on the murmur task was determined by the weighted accuracy score. The clinical outcome task on the other hand, used clinical cost score.

| Model | Metric | Training | Validation | Test | Rank |
|---|---|---|---|---|---|
| Murmur | W. acc.* | $0.497 \pm 0.083$ | 0.585 | 0.593 | 30th |
| | Cost metric | $13158 \pm 1283$ | 8866 | 13134 | |
| | Accuracy | $0.446 \pm 0.070$ | 0.423 | 0.497 | |
| | F measure | $0.403 \pm 0.055$ | 0.384 | 0.398 | |
| Outcome | W. acc.* | $0.713 \pm 0.042$ | 0.732 | 0.703 | 14th |
| | Cost metric | $12315 \pm 903$ | 8720 | 12419 | |
| | Accuracy | $0.510 \pm 0.047$ | 0.537 | 0.537 | |
| | F measure | $0.465 \pm 0.061$ | 0.530 | 0.503 | |

Table 1. Scores obtained by the murmur and clinical outcome classifier on the training set (5-fold cross-validation) and the hidden validation and test set. The ranking is based on the performance on the test set.
* Weighted accuracy

## 4. Discussion and Conclusion

The challenge scores achieved by the clinical outcome model on the test set were consistent with the performance on the training set. However, the performance on the validation set was significantly better and even outperformed all other participants' models in the challenge. It should be noted that the validation set was relatively small compared to the training and test sets. This could cause a higher variance in performance by chance. Since each team got 10 attempts to test their algorithm on the validation set, we might have indirectly overfitted the models to a data set that is not representing the variation in the training and test set. This emphasizes the importance of having a large enough, diverse and completely untouched data set when testing and reporting results from machine learning studies

Pre-training models were also tested using the 2016 PhysioNet Challenge data set [14, 15]. Different approaches on training the pre-trained models were explored. However, there were no significant improvements during CV on the public training set and the performance on the validation set was actually lower compared to no pre-training.

Both murmur and the clinical outcome classifiers were trained using single PCG recordings, and the auscultation location was not taken into consideration. However, in the preliminary phase of the challenge, we also tested multi-channel PCG classifiers, but they were outperformed by the single-channel classifiers. This observation taken into account in addition to the performance of our classifier and other challenge participants' classifiers, supports the hypothesis that a CNN can detect abnormalities from PCG recordings regardless of the auscultation location. This finding add to the development of CNN-based PCG classifiers. However, further studies are needed to provide an in-depth explanation of how these CNNs interpret the PCGs. A greater focus on the explainability of these models may produce interesting findings that could be of clinical relevance.

## References


[1] WHO Cardiovascular Diseases (CVDs). URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 21.08.2022.

[2] Montinari MR, Minelli S. The First 200 Years of Cardiac Auscultation and Future Perspectives. Journal of Multidisciplinary Healthcare March 2019;12:183–189, DOI: https://doi.org/10.2147/JMDH.S193904.

[3] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022. preprint, August 2022. DOI: https://doi.org/10.1101/2022.08.11.22278688.

[4] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. IEEE Journal of Biomedical and Health Informatics June 2022;26(6):2524–2535, DOI: https://doi.org/10.1109/JBHI.2021.3137048.

[5] Vukanovic-Criley JM, Criley S, Warde CM, Boker JR, Guevara-Matheus L, Churchill WH, et al. Competency in Cardiac Examination Skills in Medical Students, Trainees, Physicians, and Faculty: A Multicenter Study. Archives of Internal Medicine March 2006;166(6):610–616, DOI: https://doi.org/10.1001/archinte.166.6.610.

[6] Roy D, Sargeant J, Gray J, Hoyt B, Allen M, Fleming M. Helping Family Physicians Improve their Cardiac Auscultation Skills with an Interactive CD-ROM. The Journal of Continuing Education in the Health Professions 2002;22(3):152–159, DOI: https://doi.org/10.1002/chp.1340220304.

[7] Barrett MJ, Mackie AS, Finley JP. Cardiac Auscultation in the Modern Era: Premature Requiem or Phoenix Rising? Cardiology in Review October 2017;25(5):205–210, DOI: https://doi.org/10.1097/CRD.0000000000000145.

[8] Rangayyan RM, Lehner RJ. Phonocardiogram Signal Analysis: A Review. Critical Reviews in Biomedical Engineering 1987;15(3):211–236.

[9] Emmanuel BS. A Review of Signal Processing Techniques for Heart Sound Analysis in Clinical Diagnosis. Journal of Medical Engineering Technology August 2012;36(6):303–307, DOI: https://doi.org/10.3109/03091902.2012.684831.

[10] Debbal SM, Bereksi-Reguig F. Computerized Heart Sounds Analysis. Computers in Biology and Medicine February 2008;38(2):263–280, DOI: https://doi.org/10.1016/j.compbiomed.2007.09.006.

[11] Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. Deep Learning Methods for Heart Sounds Classification: A Systematic Review. Entropy Basel Switzerland May 2021;23(6):667, DOI: https://doi.org/10.3390/e23060667.

[12] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation June 2000;101(23), DOI: https://doi.org/10.1161/01.CIR.101.23.e215.

[13] Ismail Fawaz et al. H. InceptionTime: Finding AlexNet for Time Series Classification. Data Mining and Knowledge Discovery November 2020;34(6):1936–1962, DOI: https://doi.org/10.1007/s10618-020-00710-y.

[14] Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. Classification of Normal/Abnormal Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In 2016 Computing in Cardiology Conference (CinC). September 2016; 609–612.

[15] Clifford GD, Liu C, Moody B, Millet J, Schmidt S, Li Q, et al. Recent Advances in Heart Sound Analysis. Physiological Measurement August 2017;38(8):E10–E25, DOI: https://doi.org/10.1088/1361-6579/aa7ec8.



Address for correspondence:

Bjørn-Jostein Singstad
Kristian Augustus Gate 23, 0164 Oslo, Norway
bjornjs@simula.no