Murmur Identification Using Supervised Contrastive Learning

Ľubomír Antoni[2], Erik Bruoth[2], Peter Bugata[1], Peter Bugata Jr.[1], Dávid Gajdoš[1], Dávid Hudák[1], Vladimíra Kmečová[1], Monika Staňková[1], Alexander Szabari[2], Gabriela Vozáriková[2]

[1]VSL Software, a.s., Košice, Slovakia
[2]Pavol Jozef Šafárik University in Košice, Košice, Slovakia

Abstract

As part of the George B. Moody PhysioNet Challenge 2022, we developed a computational approach to identify abnormal cardiac function from phonocardiograms that combines deep learning and traditional machine learning methods. We adopted a supervised contrastive learning and a deep convolutional neural network to obtain an embedding of the phonocardiogram slice onto a unit hypersphere in low-dimensional space. Thus, we applied the obtained latent factors to classify patients using a Random Forest model. The murmur detection classifier created by our team CeZIS received a weighted accuracy score of 0.796 (ranked 3^{rd} out of 305 submissions) and Challenge cost score of 9479 (ranked 22^{nd} out of 305 submissions) on the hidden validation set.

1. Introduction

Non-invasive evaluation of the mechanical function of the heart using a digital stethoscope and automatic evaluation of the phonocardiogram (PCG) can provide early information about congenital and acquired heart diseases in children. The George B. Moody PhysioNet Challenge 2022 [1] set the task of classifying patients according to the presence of murmurs and overall clinical outcomes based on PCGs collected from multiple auscultation locations.

2. Methods

For the classification of PCG recordings by the traditional machine learning methods, a set of the appropriate features is usually described with the help of medical experts. The evaluation of these features often requires segmentation of the PCG or its transformation to the frequency domain. In our solution, we work only with the raw unsegmented signal, from which we automatically extract latent factors for each PCG with a particular neural network. Subsequently, the obtained latent factors can be used by traditional machine learning methods to classify patients.

2.1. Training Data

For training, we use only data from the Challenge training set, which contained 3163 PCGs from 942 patients [2]. Basic demographic data (gender, age group, height, weight, pregnancy status) and at least one recording from at least one prominent auscultation location were available for each patient. Four standard locations were used: pulmonary valve (PV), aortic valve (AV), mitral valve (MV), and tricuspid valve (TV). From the point of view of the presence of murmur, patients were assigned into three classes: Absent (695 / 73.78%), Present (179 / 19.00%), and Unknown (68 / 7.22%).

Regarding the original labels of patients and their auscultation locations with observed murmur (attribute Murmur locations in dataset), we proposed our own Murmur label for each PCG (see Table 1). In more detail, the PCGs of each patient with a murmur present are divided into two subclasses according to whether a murmur was detected on the PCG at the respective location.

	Murmur label on PCG	PCG count	
A	Absent	2391	75.59%
P1	Present at current location	499	15.78%
P2	Present only at other location	117	3.70%
U	Unknown	156	4.93%

Table 1. Our proposed labeling of PCGs.

According to clinical outcome diagnosed by a medical expert, the patients were divided into two classes: Abnormal (456 / 48.41%) and Normal (486 / 51.59%). In most cases, patients with the Murmur label of Present are assigned to the Outcome label of Abnormal. However, there are 29 patients in the training set with the observed murmur and the Outcome label of Normal. The number of patients with the Outcome label of Abnormal is much higher than patients with the observed murmur. Surprisingly, it corresponds to almost half of a number of the children who participated in the screening.

2.2. Contrastive Learning

Recently, contrastive learning (e.g., SimCLR framework [3]) has become popular in the field of supervised representation learning. Two stochastic data augmentations are used to transform each sample within a batch, yielding two correlated views of that sample. Thus, with the help of a particular contrastive loss function, an embedding on the unit hypersphere is found, which ensures that each selected sample from the batch (anchor) is close to the paired view in the embedding space and, conversely, is far from the other sample views in the batch.

A similar concept can be fruitfully applied in a fully supervised learning scenario where labels can be used. The SupCon loss function [4] ensures that for each selected anchor, other samples from the same class in the batch are located in the embedded space nearby, while samples from others classes are located much further away (Figure 1). The obtained embedding can be used to solve the original classification task, but also for downstream tasks.

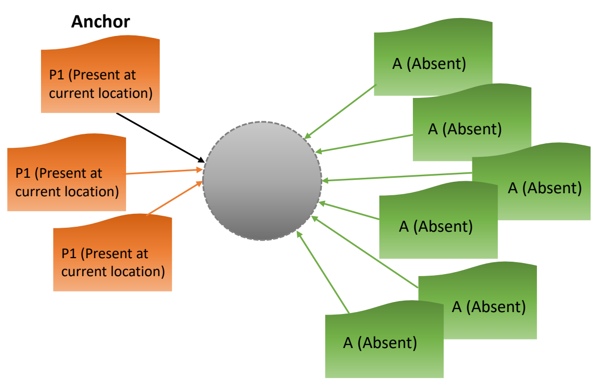

Figure 1. The concept of supervised contrastive learning.

By adapting the concept of supervised contrastive learning (SupCLR) for PCGs, we obtain an embedding for a short PCG (approximately 8 seconds) onto the unit hypersphere in a low-dimensional space (dim = 16). The input is a raw PCG signal resampled to 1000 Hz with minimal preprocessing further processed by a deep convolutional neural network (CNN). The backbone of the CNN is a one-dimensional variant of the ResNet50 network [5] with a reduced number of virtual channels (width $= 1/4$), to the output of which L_2 normalization is applied. Thus, for each PCG slice, we obtain its embedding onto the unit hypersphere in 512-dimensional space. Subsequently, a projection head is added, which maps the 512-dimensional space to the 16-dimensional space. In our solution, the projection head contains only one simple linear layer without hidden layers and non-linear activations. The L_2 normalization is then used again to embed the PCG slice onto the unit hypersphere in 16-dimensional space. The architecture of the solution is shown in Figure 2.

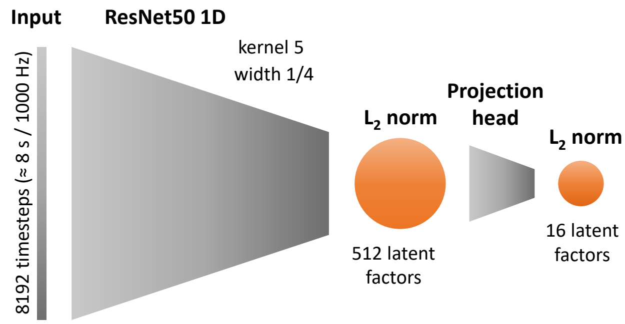

Figure 2. Neural network architecture of our solution.

For contrastive learning, only PCGs with murmur labels A (Absent) and P1 (Present/Murmur location) described in Section 2.1 are used. The label P2 is omitted due to its higher difficulty of the correct classification in comparison with the label A. Moreover, for the patient labeled Unknown, it is not clear whether all his/her PCGs are technically problematic or only some of them. Thus, the label U is excluded from the contrastive learning, as well.

In this way, a binary classification task with clearly separable classes was obtained. One class includes the PCGs without murmur and the other contains the PCGs with an audible murmur. In comparison with [4], the size of the batch is not doubled during training. Due to the sufficiently large batch size (512), there are always at least two samples from the minority class in the batch. We trained the model using the AdamW optimizer with a weight decay of 0.0005 and the OneCycle learning rate [6] schedule with a maximal learning rate of 0.02 total of 100 epochs.

2.3. Patient Classification

In our solution by supervised contrastive learning described in the previous section, we constructed 16 latent factors for each PCG lasting approximately 8 seconds. These factors are further used to create an additional dataset to train patient classification models.

PCGs in the training set range in length from approximately 5 seconds to more than 64 seconds. For each PCG, 10 different slices (offsets) are defined with a length of approximately 8 seconds. The offsets are evenly distributed over the entire length of the PCG with overlap. If the PCG is shorter than 8 seconds, the signal is padded with varying numbers of zeros on the left and right, such that the non-zero part is gradually shifted from left to right. For each patient, the 10 views corresponding to the offsets are used for data augmentation (Figure 3).

For each patient view, we created the explanatory variables shown in Table 2. Since the PCGs entering the CNN are standardized, the latent factors do not contain information about the mean and standard deviation (StDev) of the signal. The values of the demographic variables are

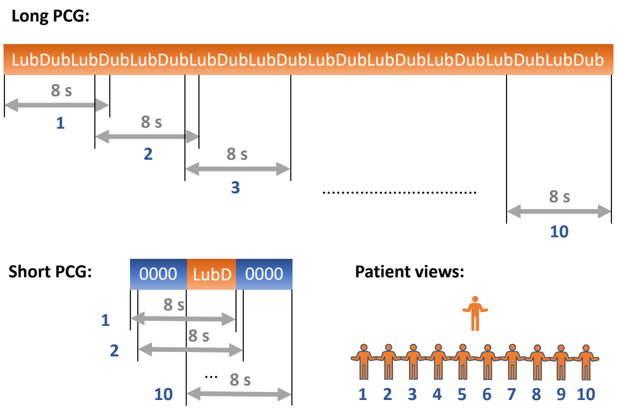

Figure 3. PCG slices for patient views.

constant for all views of the patient. However, the values of the other variables can vary for each view. Thus, we obtain a dataset with 77 explanatory variables, and the target variable Murmur with three classes (Absent, Present, Unknown). We multiplied the number of dataset rows 10 times the number of patients. Thus, we obtained more samples for training to prevent overfitting. We also applied views for predicting a patient label by the arithmetic mean of predicted probabilities from all patient views.

Explanatory variables	**Locations**	**Count**
16 latent factors from SupCLR	4	64
Mean and StDev of the PCG slice	4	8
Patient demographic data	-	5

Table 2. Explanatory variables for patient classification.

We did not solve the given classification task as multi-class task. However, we divided the task into two subsequent binary classifications (Figure 4):

1. **Present** vs. **others (Absent, Unknown)** – to separate the patients with the original Murmur label of Present.
2. **Unknown** vs. **Absent** – the additional classification for other values of the original Murmur labels of the patients.

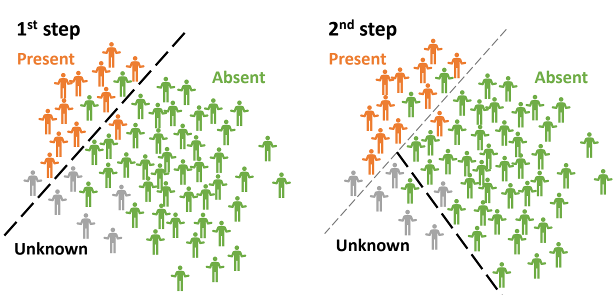

Figure 4. Subsequent classification steps.

We selected the well-known and popular Random Forest (RF) method to solve the classification tasks. We applied its "balanced" version [7] implemented in the python library imbalanced-learn [8], which allows setting the number of randomly selected samples for each class. By a grid search method, we found the optimal values of the following hyperparameters for each classification:

- The ratio of samples selected from the positive and negative classes for building the tree;
- Probability threshold for positive class prediction.

With these hyperparameters, it is possible to influence the number of predictions into the positive class so that the highest possible value of the metric (weighted accuracy in this case) is achieved.

For the Outcome label, PCGs are divided into two classes based on whether the patient's label is Abnormal or Normal. Using these labels on PCG along with supervised contrastive learning, 16 new latent factors are obtained. For patient classification, an auxiliary dataset is created similarly as for the Murmur label, it differs only in 16 variables corresponding to new latent factors. For this binary classification, only one balanced RF is created. By the grid search, the values of the hyperparameters were chosen to get the lowest possible value of the special metric defined as the average price for the diagnosis and treatment of one patient using the given model for patient pre-screening.

3. Results

The quality of the trained models was evaluated using 10-fold cross-validation (CV) on the training set. We divided the patients into 10 folds stratified by both target variables: Murmur and Outcome. For evaluation objectivity, we ensured that all PCGs of the same patient were in the same fold, even if he/she participated in both campaigns and was assigned two patient identifiers.

For each fold from the 10-fold CV, a model including one CNN and one or two RFs was created. For prediction on the validation and test set, an ensemble model consisting of all these 10 models was used to obtain the final prediction by applying their voting. Thus, one PCG from each standard location was considered for patient label prediction, while 10 offsets were considered for each PCG, and the prediction was obtained using 10 CNNs for each offset.

3.1. Murmur Label

To evaluate the Murmur label prediction, a weighted accuracy metric was used, with the weights 5, 3, and 1 for the classes Present, Unknown, and Absent, respectively. The values achieved on the training, validation, and test set are shown in Table 3. For the training set, the mean and standard deviation obtained from CV are shown, and for the test set, the ranking among all teams is also displayed.

A more detailed study of the confusion matrix indicates that the created model has the highest error rate when classifying PCGs from the Unknown class.

Training	Validation	Test	Ranking
0.804 ± 0.037	0.796		

Table 3. Results for Murmur label.

3.2. Outcome Label

The metric for the Outcome label was defined as a price that includes the costs of algorithmic pre-screening, expert screening, treatment, and diagnostic errors that result in late treatments. The achieved results are shown in Table 4.

Training	Validation	Test	Ranking
10984 ± 1067	9479		

Table 4. Results for Outcome label.

The analysis of the confusion matrix in this binary classification shows a high error rate of the model. Due to the inaccuracy of the model and the high cost of late treatment, the model predicts a very large proportion of patients as Abnormal and refers them to experts for examination.

The large difference in the results achieved on the validation and training set indicates that the validation set is relatively small and may not represent the entire dataset quite accurately.

4. Discussion and Conclusions

The applications of contrastive learning in the health care area can be beneficial in several ways. In the case of a sufficient amount of data, but for which labels from medical experts are not available, self-supervised contrastive learning can be used to obtain an appropriate data representation. Our presented solution indicates that contrastive learning can be successfully applied in a fully-supervised setting even with a smaller amount of data, as well.

In the first task, it was possible to create models that distinguish the presence and absence of murmur quite well. Since the identification of the Unknown class seems problematic, we see the potential for algorithm improvement in the addition of a special evaluation of the technical quality of the recording before the murmur identification.

The second task of predicting the clinical outcome appears to be more demanding and the created models are very inaccurate. The difficulty of the task lies in the fact that for patients marked as Abnormal, it is not clear on which of their PCGs (if any) the anomaly identified during the comprehensive examination manifests itself. In future, it is possible to consider a one-class classification trained only on patients marked as Normal.

Acknowledgments

This work was supported by ERDF EU grant and by the Slovak Ministry of Economy under the contract No. ITMS313012S703 and by the Slovak Grant Agency for Science (VEGA) under contract 1/0177/21.

References

[1] 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. medRxiv 2022;URL `https://doi.org/10.1101/2022.08.11.22278688`.

[2] 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 2021;26(6):2524–2535.

[3] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, ICML'20. JMLR.org, 2020; 1597–1607.

[4] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds.), Advances in Neural Information Processing Systems, volume 33. Curran Associates, Inc., 2020; 18661–18673. URL `https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf`.

[5] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[6] Smith LN, Topin N. Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 2019; .

[7] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, UC Berkley, 2004. URL `http://xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html`.

[8] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 2017;18(17):1–5. URL `http://jmlr.org/papers/v18/16-365.html`.

Address for correspondence:

Dávid Hudák
VSL Software, a.s., Lomená 8, 040 01 Košice, Slovakia
hudak@vsl.sk