2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Classification of real-world pathological phonocardiograms through multi-instance learning

Andrea Duggento[1], Allegra Conti[1], Maria Guerrisi[1], Nicola Toschi[1,2]

***Abstract*— Heart auscultation is an inexpensive and fundamental technique to effectively to diagnose cardiovascular disease. However, due to relatively high human error rates even when auscultation is performed by an experienced physician, and due to the not universal availability of qualified personnel e.g. in developing countries, a large body of research is attempting to develop automated, computational tools for detecting abnormalities in heart sounds. The large heterogeneity of achievable data quality and devices, the variety o possible heart pathologies, and a generally poor signal-to-noise ratio make this problem extremely challenging. We present an accurate classification strategy for diagnosing heart sounds based on 1) automatic heart phase segmentation, 2) state-of-the art filters drawn from the filed of speech synthesis (mel-frequency cepstral representation), and 3) an ad-hoc multi-branch, multi-instance artificial neural network based on convolutional layers and fully connected neuronal ensembles which separately learns from each heart phase, hence leveraging their different physiological significance. We demonstrate that it is possible to train our architecture to reach very high performances, e.g. an AUC of 0.87 or a sensitivity of 0.97. Our machine-learning-based tool could be employed for heart sound classification, especially as a screening tool in a variety of situations including telemedicine applications.**

I. Introduction

With more than 18 million deaths per year, cardiovascular disease is both the leading cause of death and disability worldwide.The problem is worse in developing countries, where lack of medical professionals prevents or significantly hampers early detection of the disease. To this end, with the exception of diseases confined to brain vasculature, or other peripheral conditions (such as deep vein thrombosis), the majority of cardiovascular diseases – such as coronary heart disease, rheumatic heart disease, congenital heart disease, heart valve disease – can in principle be diagnosed by trained physicians using only inexpensive equipment. Classically, physicians are trained to recognise few fundamental heart sounds (FHSs) which are produced by mechanical phenomena inherently connected to heart anatomy and physiology, such as closure of a valve or tensing of a chordae tendineae [1]. Among the most conspicuous and hence diagnostically useful FHSs are the first (S1) and second (S2) heart sounds. In Fig. 1 an example of typical phonocardiogram (PCG) is shown.

With the advent of Machine Learning (ML) applications in medicine, in the last 25 years various dedicated computer-aided diagnosis (CAD) systems for heart diseases have been proposed [3]. In general, CAD system are designed and trained for highly detection sensitivity – which typically corresponds to low specificity – and are therefore employed in mass screening designed to highlight suspected pathological cases to a trained physician. This priority however does not apply when the goal is the development of a fully automated system. The analysis pipeline of the vast majority of all the CAD systems that has been proposed for PCG discrimination is composed of two main steps: 1) preprocessing (denoising and segmentation of the FHSs) 2) feature extraction and classification. A number of techniques have been proposed for both step 1) and 2) (see [4] for a general overview). In spite of the multitude of attempts, the performances provided by these methods vary greatly, and, importantly, are typically specific to a certain heart condition, hence severely hampering generalizability to the clinical context. Also, as pointed out in [5], a number of these studies might be flawed in several ways, including the lack a separate test set, ill-documented data, a-priori exclusion on noisy data, or creation of overly homogeneous datasets not representative of real-life conditions.

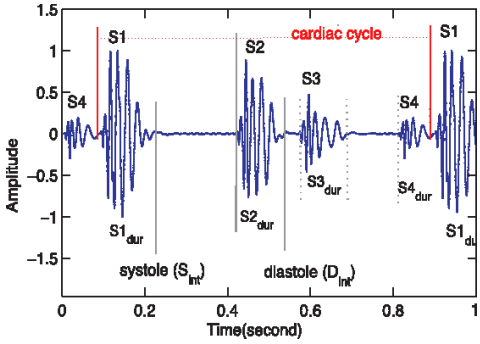

Fig. 1. Example of a typical normal PCG signal where the different heart sounds (S1, S2, S3, S4) are labelled with reference to the cardiac cycle [2].

The aim of this paper was to devise a classification pipeline which would be as accurate as expert clinician diagnosis while also remaining generalizable to multiple pathologies, robust to low recording quality and differences in recording lengths while exploiting expert knowledge about the significance and meaning of different heart sound outlined above.

II. Methods

In the following we illustrate our pipeline composed by 1) automatic heart phase segmentation which employ a hidden Markov model, 2) MFCC representation of segmented heart

[1]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy
[2]Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Boston, MA, USA

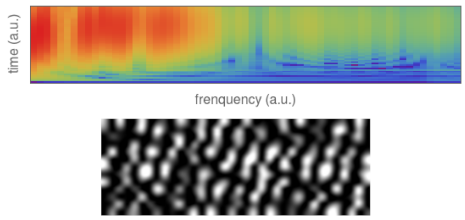

Fig. 2. Top: Example of Mel spectrogram as derived from a 300 ms long diastolic-phase sound segment. Bottom: representation of MFCC decomposition on a scale from 0 (black) to 1 (white).

sounds), and 3) layers of CNN followed by fully connected neuronal ensembles in a multi-instance learning architecture which we constructed in order to leverage the presence of different types of information (i.e. sounds) originating from a single example (i.e. subject) with unknown condition(s).

A. Data

The data employed in this paper consisted of muticentric database of heart sound recordings (HSR) which is described in detail in [5]. It is part of a public challenge for heart sound classification called 'The PhysioNet Computing in Cardiology Challenge 2016' was launched [6] and the related heart sound database was made publicly available on PhysioNet [7]. It was designed to contain noisy, multicentric, multipathology recordings of very different lengths, recorded at a number of difference centres (i.e. a real-world database). It collects nine different databases, recorded over more than a decade, by seven teams in seven different countries and three continents, from a variety of clinical or nonclinical (such as in-home visits) environments and equipment. Given the heterogeneity of the databases, data quality and patient types differ considerably, and the only information provided is the class assigned by an expert physician: "normal" or "abnormal", without any indication about e.g. etiology, pathology, comorbidities or any patient demographics at all. The recording length varied from a few seconds seconds to a few minutes.

We employed 3153 HSR, drawn from 764 subjects and patients with different conditions, including heart valve and coronary artery disease, for a total of 84425 beats. While the recording sample rate at acquisition varied from 800 Hz up to 44,100 Hz (with 4000 Hz being the most common value), all HSR were resampled and archived in .wav format at 2,000 Hz sampling frequency.

B. Segmentation of heart sounds

For segmenting the PCG recordings into four different non overlapping heart phases (S1, S2, systolic phase, diastolic phase), in this paper a duration-dependent hidden Markov model (DHMM) [8] was employed. The DHMM was previously validated in [8] trained with manually labelled recordings acquired using a commercially available electronic stethoscope, with a large variety of the signal-to-noise ratios, and included patients with valvular heart disease, arrhythmia, pulmonary diseases and obesity, and the recordings were contaminated with both physiological noise and background noise. Given that segmentation results in samples of different length (between subjects as well as between heart beats), a minimum phase-dependent time length was imposed by discarding: 100 ms for S1, 180 ms for systole, 100 ms for S2, 300 ms for diastole.

C. Mel spectrogram and MFCC

For each phase, the data was mapped onto a Mel spectrogram space [9]. To obtain the MFCC, the Fourier discrete cosine transform of the logarithm of each frame of the mel-spectrogram was computed (see Fig. 2), thus dramatically reducing the dimensionality of the features.

D. Multi-branch multiple-instance NN

Each subject results in a set of sounds for each heart phases whose size varies in length between subjects. In view of this, a multi-branch multiple-instance network was trained to discriminate 'normal' heart recordings from 'abnormal' heart recordings from four, non necessarily contiguous sounds, (one sound for each phase).

Multiple instance learning (MIL) is considered a weakly supervised method since the training data elements are pooled into so-called 'bags'. No label is required for each single element within each bag, and the supervision is required only at the 'bag-level', i.e. a label is provided for the bag. Recently this approach has gained a lot of attention because of its ability to labels large dataset in multiple science fields.

In detail, we employed an artificial neural network (ANN) branch for each of the four heart phases. All four branches were composed of several convolutional layers with rectified linear unit (ReLU) activations and max-pooling layers. As the dimensions of the MFCC were different across heart sound segments, the exact numbers of convolutional layers was branch-specific, as detailed on Fig. 3. In all of the four branches the CNN was piped to a fully connected layer of two neurons with sigmoid activations. Finally, all neurons on the last layers from the for branches were fully connected to the last layer of sigmoidal activated neurons which provide final discrimination.

To train the architecture, data was fed to it in the following way: each recordings was first segmented with the DHMM method, obtaining four 'bags' of segments, each corresponding to a heart phase. Multiple recording from the same patient were allowed into the training set. A random choice of one segment per-'bag' was extracted. Thus extracted segments were typically not contiguous in time. This choice represented a single-instance and was labelled 'normal' or 'abnormal' accordingly to the patient label. Each single segment was then analysed to extract the MFCC representation, and fed to the multi-branch CNN architecture. Collecting a (four-segment) instance from each recording represents a step, and 3500 optimization steps where repeated at each epoch before the gradient-descent controlled update of the ANN parameters.

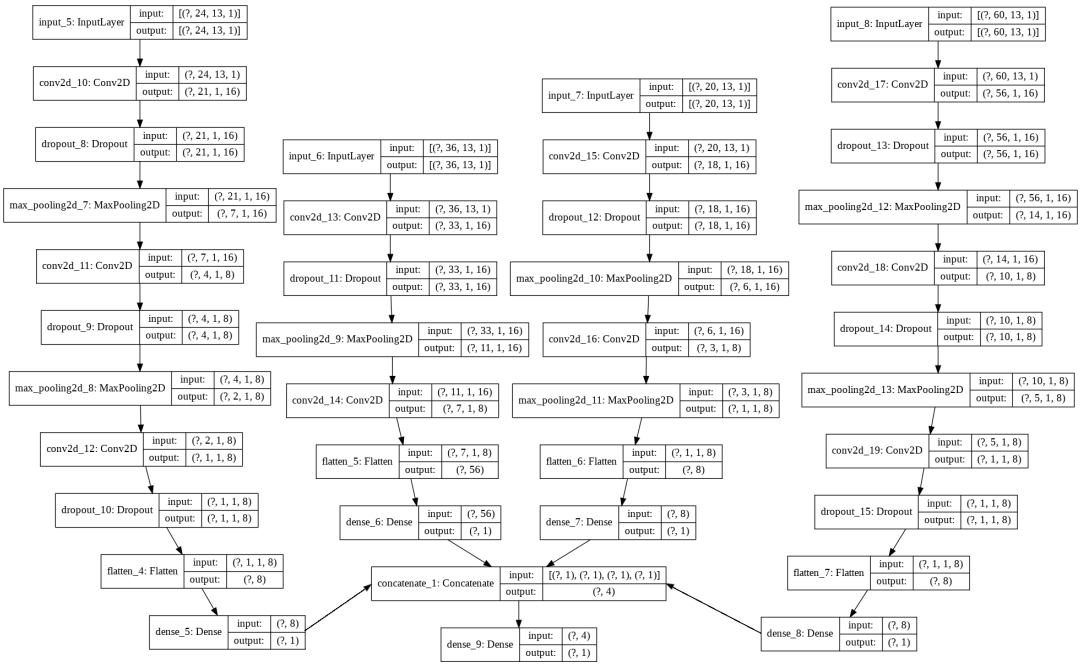

Fig. 3. Overview of the whole architecture of the multi-branch multi-instance neural network architecture.

As commonly done, in order to determine the optimal number of training epochs to maximize training quality while avoiding overfitting, the dataset was split into train/validation sets with a 80/20 ratio. The optimal number of epochs is determined empirically as the one where overfitting is about to begin (i.e. increasing performance on the training but not on the test set). Successively, we retrained the model in a 95/5 leave-p-out fashion (performed 128 times through random sampling with replacement) framework for performance estimation. Performance was quantified through i) accuracy (ACC), ii) specificity, iii) precision/positive predictive value (PPV), iv) sensitivity/true positive rate (TPR). The operating point is typically chosen depending on the clinical needs, i.e. whether sensitivity and specificity are equally important, or rather if one of the two should be privileged. Commonly, The optimal working threshold is chosen by maximizing the harmonic mean of the PPV and TPR, with additional weights if precision and recall have different clinical importance. For this purpose, the F_β score is introduced:

$$F_\beta = (1+\beta^2)\frac{PPV \times TPR}{(\beta^2)PPV + TPR}$$

where β is commonly chosen between three values: $\beta = 1$ (unweighted harmonic mean PPV and TPR), $\beta = .5$ (which weighs recall less than precision), and $\beta = 2$ (which weighs recall more than precision).

III. Results

In Fig. 4 (left) the sample accuracy with respect to training epochs is shown. ON the left it can be clearly seen that after a steep learning for about 60 epochs, overfitting comes into play (increasing accuracy on the training set and a virtually constant accuracy on the test set). To obtain confidence

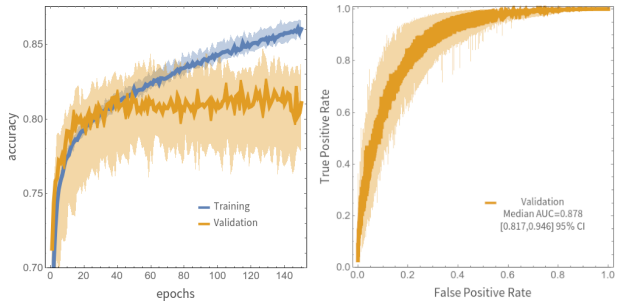

Fig. 4. Left: sample accuracy with respect to training epochs for an 80/20 train/validation split. The median accuracy across splits for the training test sets is shown as solid lines, along with a $\pm$ 2 standard deviation interval (shaded areas). Right: ROC curves obtained in test data for $2^7 = 128$ different 95/5 training/validation splits (random sampling with replacement)

intervals for the accuracy, the procedure was repeated 10 times on 10 different of training/validation splits.

To obtain a figure of merit in terms of accuracy, specificity, precision and the sensitivity, the ratio of training/validation ratio was set to 95%/5%. To account higher variability on the smaller test set, $2^7 = 128$ different splits were considered. The training lenght was *a priori* set to 60 epochs (3500 training step each). Median ROC curve with 95% confidence intervals on test sets is shown on Fig. 4 (right). Median area under ROC curve (AUC) across splits in test sets was 0.878 $\pm$(0.817,0.946) 95% CI.

For every split we evaluated the accuracy, the specificity, the precision and the sensitivity in the test test at every threshold operating points. In particular, we considered three operating points at which the f_1-, the $f_{0.5}$- and f_2-score were maximized, respectively. For each of the three threshold operating points we considered the distribution of accuracy,

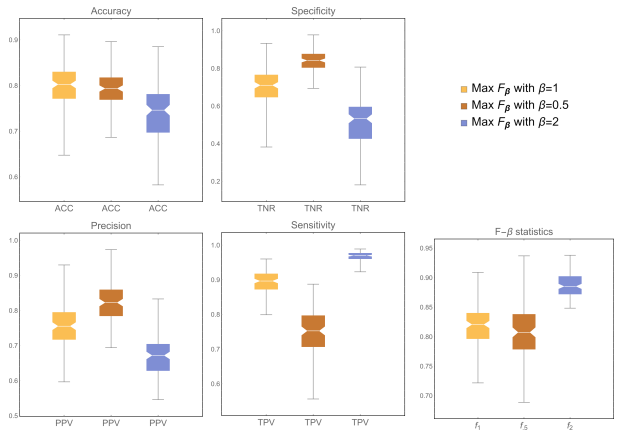

Fig. 5. Distributions of PPV and TPR and F_β scores across the 128 validation. We evaluate the distribution of In (a) and (b) sets considered Example of parameters distributions from last layer of first branch of ANN. Parameters are a function of epochs' steps.

	Max F_1 F_1=0.821	Max $F_{.5}$ $F_{.5}$=0.807	Max F_2 F_2=0.862
Accuracy	0.802	0.794	0.746
Specificity	0.71	0.842	0.532
Precision	0.755	0.822	0.672
Sensitivity	0.896	0.753	0.97

Fig. 6. Median values across $2^7 = 128$ leave-p-out ($p = 5\%$) cross-validation iterations of figures of merit for model performance: accuracy, specificity, precision, sensitivity. The figures of merit were evaluated a three different operating points which maximize F_1-, $F_{0.5}$- and F_2-score statistics respectively.

specificity, precision and sensitivity along the 128 test splits in the validation set. The box and whiskers plots on Fig. 5 show the median and the interquantile ranges of all figures of merit for performance at 3 operating points, showing extremely high performances in all settings.

On Table 6 the medians of the distributions are shown. Among the F_β-scores, F_2 reached the highest median (higher weighs in sensitivity with respect to precision).

IV. Discussion and Conclusion

The method that we proposed to tackle the problem of phonocardiogram classification is based on a pipeline that encompasses a segmentation stage of the phonocardiogram signal into four phases of the heart (S1, S2, systolic phase, diastolic phase), frequency decomposition based on the MFCC representation, and a multi-branch, multi-instance ANN that takes as input the MFCC representation of four, not necessarily contiguous, heart phases and return a single scalar value bounded between 0 and 1 (value increases with the predicted probability of a pathological recording).

We achieved extremely good results considering the heterogeneity of the database including signal to noise ratio, underlying pathologies, multicentric nature, mixed ages and demographics, comorbidities etc - all data which was not available to us. We therefore are confident that, in contrast to previous papers on this topic, our method is specifically tailored to be high performing and robust in real-world application.

The level of the classification threshold should be adjusted within the context of the clinical needs. For instance, in applications where false positive cases lead to greater medical and/or financial burden with respect to the consequences of false negative cases, (i.e. specificity is more important than sensitivity) a higher threshold is preferred, and usually a conservative statistics such as $F_0.5$-score is considered. Conversely, when high detection sensitivity is needed, and high number of false positive cases lead to minor issues with respect to possible false negative cases, a more 'sensitive' statistics such as F_2-score is preferred like e.g. in mass screening.

With a median F_2-score higher that both F_1- and $F_{0.5}$-scores, our results indicate that the proposed method is better suited as a first-line screening application. Indeed, at highest F_2-score the median sensitivity is higher than 97% (but at the expenses of a relatively low specificity). In this context, it should be noted that such figure of merit is averaged with a leave-p-out cross-validation method due to the unavailability of a dedicated test set, which is privately retained by organizers of the PhysioNet computing in cardiology challenge 2016 [6] for independent assessments.

Relatively high performances obtained in this first study suggest that the method might prove useful as e.g. screening tools in a variety of situations including telemedicine applications. Further work is necessary in order to refine the classification architecture towards specialization into heart-phase based classification, and hence more informed clinical diagnosis.

References

[1] V. Voin, R. J. Oskouian, M. Loukas, and R. S. Tubbs, "Auscultation of the heart," *Clinical Anatomy*, vol. 30, no. 1, pp. 58–60, sep 2016.

[2] V. N. Varghees and K. Ramachandran, "A novel heart sound activity detection framework for automated heart sound analysis," *Biomedical Signal Processing and Control*, vol. 13, pp. 174–188, sep 2014.

[3] L. Bahekar, A. Misal, and G. Sinha, "Heart sound segmentation techniques: a survey," *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE), e-ISSN*, pp. 2278–1676, 2014.

[4] M. Nabih-Ali, E.-S. A. El-Dahshan, and A. S. Yahia, "A review of intelligent systems for heart sound signal analysis," *Journal of Medical Engineering & Technology*, vol. 41, no. 7, pp. 553–563, oct 2017.

[5] C. Liu and et Al., "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.

[6] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *2016 Computing in cardiology conference (CinC)*. IEEE, 2016, pp. 609–612.

[7] A. L. Goldberger and et Al., "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[8] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden markov model," *Physiological measurement*, vol. 31, no. 4, p. 513, 2010.

[9] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, apr 2019.