Towards Uncertainty-Aware Murmur Detection in Heart Sounds via Tandem Learning

Erika Bondareva, Tong Xia, Jing Han, Cecilia Mascolo

University of Cambridge, UK

Abstract

Auscultation, the process of using a stethoscope for diagnostics, is a challenging task for medical professionals and requires years of training. As a result, the field of automated auscultation has been growing in popularity in the past decade. Previous efforts in the field focused on achieving high accuracy, with confident, albeit sometimes wrong, classifiers. Such model over-confidence is especially dangerous in healthcare setting. Leveraging the release of the new heart sound dataset as a part of PhysioNet 2022 challenge, we explored a novel murmur detection methodology using uncertainty-aware tandem learning. In order to effectively separate unknown samples and detect heart sounds with murmur present, we developed two binary classifiers, under the assumption that training two models to solve simpler tasks could improve the overall sensitivity. First, a support vector machine used spectral features for identification of unknown samples. We then used a Deep Neural Network (DNN) with a set of hand-crafted audio features for prediction of murmur. In addition, we implemented uncertainty estimation in DNN using Monte Carlo dropouts for further eliminating any samples that should be labelled as unknown. With our approach we achieved 63% and 69% sensitivity and specificity of murmur, scoring 0.519 and 11301 during the challenge for murmur and outcome prediction tasks, respectively.

1. Introduction

Automated cardiac auscultation could promote preventative healthcare and improve the standard of care: manual auscultation is a challenging task for clinicians and requires years of training.

With the emergence of deep learning, researchers in academia and industry are actively exploring novel applications. The main difficulty in achieving deep learning's potential in healthcare is a clear lack of high-quality large datasets. Among body sounds datasets, the field of machine learning for heart sounds (HSs) is among the most mature, but even for HSs the vast majority of previous work is based on only two datasets, released as part of challenge: The PASCAL Classifying Heart Sounds Challenge 2011 [1] and The PhysioNet/Computing in Cardiology Challenge 2016 [2]. However, both of these datasets focus solely on binary classification of HSs. An additional limitation is that the HS samples from the same patient are treated as separate, independent samples, as if they were from different patients. Given that the murmur intensity could vary for various auscultation locations for a single patient, this misses the opportunity to use multiple sounds from a single patient for a more accurate diagnosis.

PhysioNet released a new dataset as a part of Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022 [3], which, for the first time, addresses some of the issues with the existing HS datasets. Firstly, it goes beyond binary labels and introduces an *unknown* label. It also includes a detailed description of the type of the murmur in the metadata, allowing for a more detailed analysis of the abnormal HS. Secondly, multiple HS recordings from different auscultatory locations are available for a single patient, an opportunity to leverage this data for more precise diagnostics.

PhysioNet 2022 proposed two tasks for challenge participants: first, the teams were invited to develop a classification algorithm for three classes: murmur present (further referred to as *murmur*), murmur absent (further referred to as *normal*), and *unknown*. The second task revolved around getting the lowest outcome score possible, where the *abnormal* class included both *murmur* and *unknown* sounds, while the *normal* class included the sounds with the murmur absent.

Unknown samples could in reality be either *normal* or *murmur* samples, and the misclassification of *unknown* samples as *normal* was heavily penalised by the score system. In addition, distinguishing *murmur* samples from *normal* is a challenging task, and a classification system that could flag particularly difficult to classify cases as *unknown*, as well as provide the uncertainty for the prediction, could significantly reduce the risk of misdiagnosis. We developed a tandem learning approach to leverage these aspects. Specifically, we first deployed a binary support vector machine (SVM) classifier to distinguish *un-*

known samples from all other samples. Subsequently, we extracted large-scale handcrafted features and fed them into a Deep Neural Network (DNN), trained to differentiate between *murmur* and *normal* samples. Our DNN had a built-in uncertainty awareness component that leveraged Monte-Carlo dropouts: this allowed us to alter the prediction for high uncertainty samples back to *unknown*. With this tandem learning strategy, we decomposed the original complex task into two simpler ones and thus lowered the risk of misclassifying the *unknown* samples. We also explored the benefit of ensemble approach for this task.

The main contributions of this paper are as follows:

- We describe a tandem learning pipeline for heart sound classification, which achieves a sensitivity of up to 63% and specificity of up to 69% for murmur detection;
- We demonstrate that a two-step (tandem) approach performs better than a single-step – three-class approach for *normal* and *unknown* detection;
- We implement, for the first time, uncertainty estimation via Monte Carlo dropouts, for heart sound classification.

2. Methods

2.1. Dataset description

PhysioNet 2022 dataset [4] contains heart sound labels per patient, as well as per sample: there are 695 patients with no murmur, 179 patients with a murmur present, and 68 patients labelled as unknown. In the cases where murmur is not too severe, a patient may manifest a murmur only in certain auscultatory locations, while in other locations no murmur might be present. As a result, the proportion of normal to murmur samples is even less balanced, with 2508 normal and 499 murmur samples, changing the ratio of normal to murmur from around $4:1$ for patients to $5:1$ for samples.

2.2. Tandem learning

For the heart sound classification task into *normal*, *murmur*, and *unknown* samples, we employed a tandem approach. We focused on maintaining high sensitivity by avoiding misclassifying *murmur* or *unknown* samples by splitting the task into two sub-tasks. The **first sub-task** concerned itself with distinguishing *unknown* samples from the rest (where *murmur* and *normal* samples were joined into a single "known" class). The **second sub-task** focused on correctly identifying *murmur* samples, where only *normal* and *murmur* samples were used for training. During this step, any samples where uncertainty was too high were labelled as *unknown*. Finally, the predictions for every sample were combined to form a final diagnosis prediction for the patient. The exact pipeline can be seen in Figure 1.

2.3. Tackling class imbalance

In order to tackle the class imbalance during training, we experimented with resampling techniques: naive resampling and SMOTE [5]. Based on our preliminary results, it appeared that naively upsampling the minority class while downsampling the majority class yielded the best performance. Specifically, for the first sub-task *unknown* class was upsampled so that the resulting number of samples is three times larger than the original, and then the number of *normal* and *murmur* samples was reduced to match the number of upsampled *unknown* samples. A similar approach was used for the second sub-task, except that the *murmur* class was upsampled five times.

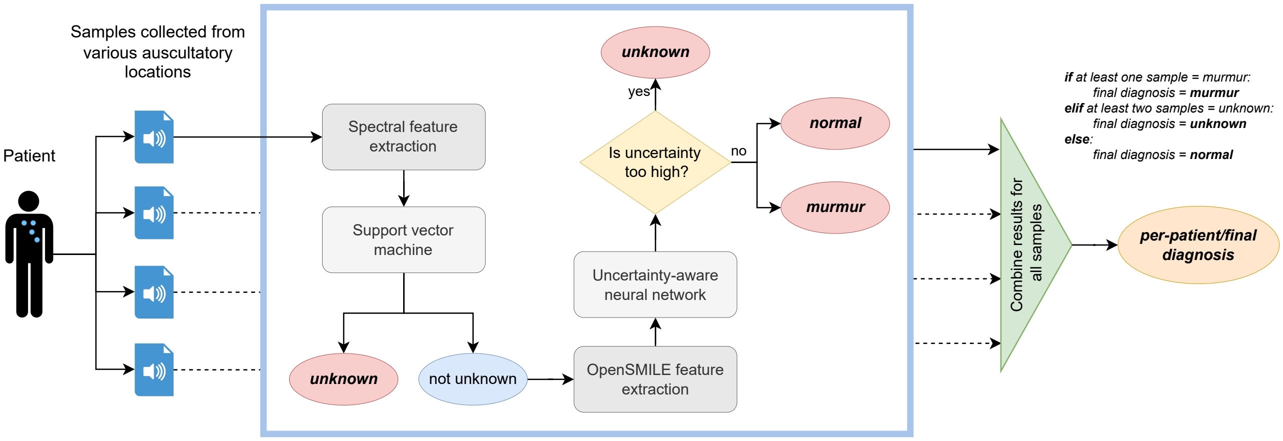
2.4. Preprocessing

According to our preliminary analysis, the best performing set of features for the first sub-task appeared to be hand-crafted spectral features, extracted over 1024 datapoints with 512 points hop length. The features extracted included chroma short time Fourier transform, melspectrogram, 40 Mel frequency cepstral coefficients (MFCCs), root mean square, spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral roll off, poly features, and, finally, zero crossing rate. For the second sub-task, we extracted the INTERSPEECH ComParE 2018 feature set (IS-18) [6], yielding 6373 features. It has been shown to perform well in a wide variety of audio-related tasks, including HS classification [7].

Moreover, for the first sub-task, after the training set was balanced by resampling, the features were scaled by removing the mean and scaling to unit variance and reduced to 0.99 variance using principal component analysis. For the second sub-task, the features were scaled but not reduced.

2.5. Prediction and evaluation

In order to detect *unknown* samples, we used an SVM with a linear kernel with hand-crafted spectral features used as inputs. Then, for *murmur* detection, we fed IS-18 features into a neural network with Monte Carlo dropout, training it for 15 epochs. The neural network consisted of 6 dense layers with relu activation and dropout of 0.5 for every layer except the first one, where the dropout was 0.2. The last layer of the network was softmax. Preserving dropout for testing, a number of predictions was obtained on the samples from the test set, and the deviation of the predictions was considered as model uncertainty. We tried various number of predictions starting from 10, but we got the best performance when obtaining 50 predictions per sample. The samples for which deviation exceeded 0.2 were then labelled as *unknown*, to further boost


Samples collected from various auscultatory locations
Patient
Spectral feature extraction
Support vector machine
unknown
not unknown
OpenSMILE feature extraction
Uncertainty-aware neural network
Is uncertainty too high?
yes
no
unknown
normal
murmur
Combine results for all samples
per-patient/final diagnosis

```
if at least one sample = murmur:
    final diagnosis = murmur
elif at least two samples = unknown:
    final diagnosis = unknown
else:
    final diagnosis = normal
```

Figure 1. A diagram demonstrating the tandem learning approach and subsequent combining of predictions per patient.

the model's sensitivity to *murmur*.

For model comparison we used two metrics as suggested by the PhysioNet challenge, namely *murmur score* and *outcome score*. We also used total *accuracy*, defined as the number of correctly classified samples divided by the total number of samples, as well as *weighted accuracy*, which is a sum of sensitivies for individual classes divided by the number of classes. In addition, we reported performance in terms of *sensitivity* and *specificity of murmur*, *sensitivity of normal*, and *sensitivity of unknown*.

For performance evaluation we implemented cross-validation, since the official validation data were not available, and the data available for training and testing the methods were limited.

In hope to further boost the performance of the algorithm for the challenge, we implemented ensemble learning. The DNN model was trained from scratch ten times for each of the validation folds, to account for performance variability induced by random weight initialisation. The three best-performing on validation set models were selected for final classification, using for evaluation the murmur score equation provided by the challenge. Majority voting scheme was exploited to derive a final prediction [8]. For unknown classification, the most sensitive classifier was chosen from 10 resulting SVMs (one SVM per fold), and the best one was used for final prediction.

Finally, to obtain the final prediction for each sample, the outcomes from both prediction stages were combined. Afterwards, to derive the final prediction for each patient, predictions for all samples belonging to the same patient were combined following a predefined rule (cf. Figure 1).

3. Experiments and Results

To evaluate the effectiveness of the proposed tandem approach, we randomly split the challenge dataset into ten patient-independent folds to carry out 10-fold cross-validation, and reported the average performance and variance across all folds. Worth noting, within each round of ten, one out of the remaining nine folds was *held out* for ensemble model selection and hyper-parameter identification, while multiple identical models were trained on the other eight folds. This process was continued iteratively until every fold out of the remaining nine was selected once as the *held out set*.

Herein, we compared our approach with a non-tandem approach. Specifically, we implemented a **three-class DNN**, where 6373 OpenSMILE features were fed into an uncertainty-aware 6-layer DNN, similar to the one deployed in the 2nd sub-task of our tandem approach, for classifying *normal*, *murmur*, and *unknown* samples in one step. Moreover, for a fair comparison, similar ensemble learning implementation was also used for this DNN-based non-tandem approach for performance boosting. As a consequence, two approaches were compared in our experiments, namely, **3-class DNN w/ ensemble**, and the proposed **tandem w/ ensemble**. The obtained performance in terms of aforementioned eight metrics for the two models are presented in Table 1.

As shown in Table 1, the tandem model outperforms the 3-class DNN model in five out of the eight metrics, yielding a mean accuracy of 0.55 cross all folds with a 14.6% relative performance improvement over the DNN model. Similarly, the obtained average performance in terms of weighted accuracy, the specificity of murmur, and the sensitivity of normal are also increased from 0.44, 0.42, and 0.42, to 0.47, 0.69, and 0.57, respectively. More importantly, we show that by leveraging two-step binary classification strategy with uncertainty score, the tandem model boosts the performance of unknown detection, leading to 0.23 for the sensitivity of unknown.

However, the tandem approach did not achieve better scores for the two official performance metrics, obtaining

Metric	**3-class DNN w/ ensemble**	**Tandem w/ ensemble**
Accuracy↑	0.48 ± 0.06	**0.55** ± **0.04**
Weighted accuracy↑	0.44 ± 0.05	**0.47** ± **0.06**
Sensitivity of murmur↑	**0.87** ± **0.09**	0.63 ± 0.14
Specificity of murmur↑	0.42 ± 0.07	**0.69** ± **0.05**
Sensitivity of normal↑	0.42 ± 0.07	**0.57** ± **0.06**
Sensitivity of unknown↑	0.03 ± 0.06	**0.23** ± **0.18**
Murmur score↑	**0.60** ± **0.06**	0.56 ± 0.05
Outcome score↓	**5769** ± **883**	6001 ± 1005

Table 1. Overall performance. Results presented are mean ± standard derivation for 10 folds.

a slightly lower murmur score and a higher outcome score when compared with the DNN model. This might be due to the much better performance of murmur sensitivity obtained by the DNN model.

Besides evaluating on data where labels were accessible, we participated in the Physionet 2022 challenge under the team name *mobihealth*, and obtained the scores 0.519 and 11301 on the official test set for murmur and outcome prediction tasks, respectively.

4. Discussion and conclusions

With the release of a new heart sound dataset as a part of Physionet 2022 Challenge, we developed a new uncertainty-aware approach for murmur detection, comparing two methodologies: two-step tandem learning and one-step three-class classification.

We observed that each of the methods studied offered unique strengths. While the one-step approach achieved better murmur and outcome scores, it suffered from a poor performance of unknown detection. The approached uncertainty-aware tandem learning, on the other hand, performed significantly better in unknown detection, and demonstrated a more balanced performance between murmur and normal detection.

It can be seen that across all methods unknown sensitivity is quite poor. This may have been caused by two issues: first, many normal samples get mislabelled as unknown; and second, while many unknown samples get classified correctly, the issues arise when combining the results per patient. Therefore, future work could focus on studying what affects the model certainty leading it to mislabel the normal samples as unknown, as well as exploring alternative methods for result combination per patient.

While the main focus of this challenge was to perform a three-class classification, the dataset's metadata contains more detailed information about the murmurs, which could be used for more granular diagnosis. Worth noting that our approach did not use segmentation, errors in which could potentially lead to misdiagnosis upon more granular classification.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L015889/1 for the Centre for Doctoral Training in Sensor Technologies and Applications, and by ERC Project 833296 (EAR).

References

[1] Bentley P, Nordehn G, Coimbra M, Mannor S. The pascal classifying heart sounds challenge 2011 (chsc2011) results. http://www.peterjbentley.com/heartchallenge/index.html.

[2] Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, Mark RG. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In Computing in Cardiology Conference. Vancouver, BC, Canada, 2016; 609–612.

[3] Reyna MA, Kiarashi Y, Elola A, et al. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. medRxiv 2022;.

[4] Oliveira J, Renna F, Costa PD, et al. The circor digiscope dataset: From murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2022; 26(6):2524–2535.

[5] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. The Journal of Artificial Intelligence Research 2002;16(1):321–357.

[6] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats. In Proc. INTERSPEECH. Hyderabad, India, 2018; 122–126.

[7] Bondareva E, Han J, Bradlow W, Mascolo C. Segmentation-free heart pathology detection using deep learning. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). 2021; 669–672.

[8] Breiman L. Bagging predictors. Machine Learning 1996; 24(2):123–140.

Address for correspondence:

eb729@cam.ac.uk