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

Erika Bondareva, Tong Xia, Jing Han, Cecilia Mascolo

University of Cambridge, Cambridge, UK

Abstract

The field of automated auscultation has been growing in popularity in the past decade due to manual auscultation being a challenging task requiring years of training. Many efforts in the field focus 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. To 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, we used a support vector machine for identification of unknown samples, followed by a Deep Neural Network (DNN) 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. Our team mobihealth achieved 63% and 69% sensitivity and specificity of murmur, scoring 0.467 (ranked 34th out of 40) and 11032 (ranked 25th out of 39) on the hidden validation set and 0.374 (ranked 40th out of 40) and 18754 (ranked 39th out of 39) on the hidden testing set 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 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.

A new dataset [3] was released as a part of Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022 [4], which, for the first time, addresses some issues with the existing HS datasets. In addition to the binary labels, it 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, presenting an opportunity to leverage this data for more precise diagnostics.

PhysioNet 2022 proposed two tasks for challenge participants, and in our work we only focused on the first task, where 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*.

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 *unknown* samples from all other samples. Subsequently, we extracted large-scale handcrafted features and fed them into a Deep Neural Network (DNN), trained to differen-

ISSN: 2325-887X DOI: 10.22489/CinC.2022.234

tiate 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 novel tandem learning pipeline for heart sound classification;
- 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 HS classification.

2. Methods

2.1. Tandem learning

For the HS 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.

2.2. 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.

2.3. 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 and averaged over the duration of the audio sample. 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].

For the first sub-task 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.

2.4. 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 the model's sensitivity to *murmur*.

For model comparison we used *murmur score*, as suggested by the PhysioNet challenge. We also used total *accuracy*, defined as the number of correctly classified samples divided by the total number of samples, *weighted accuracy*, which is a sum of sensitivities for individual classes divided by the number of classes, as well as *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 released data 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. Major-

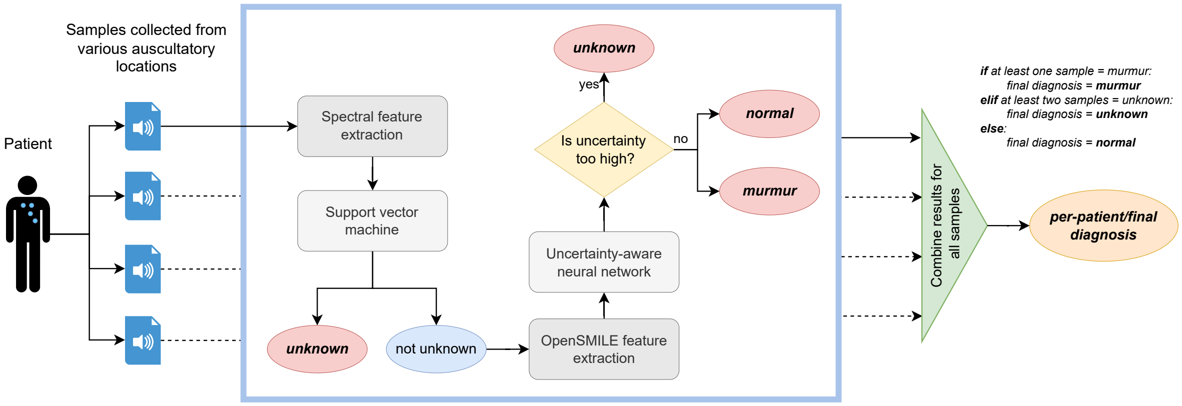
Samples collected from various auscultatory locations
Patient
Spectral feature extraction
Support vector machine
unknown
not unknown
OpenSMILE feature extraction
Uncertainty-aware neural network
Is uncertainty too high?
yes
no
unknown
normal
murmur
Combine results for all samples
per-patient/final diagnosis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

ity 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

For the proposed tandem approach evaluation, 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. 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. Therefore, two approaches were compared in our experiments: namely, the **3-class DNN w/ ensemble**, and the proposed **tandem w/ ensemble**. The obtained performance in terms of aforementioned seven metrics for the two models is presented in Table 3.

As shown in Table 3, the tandem model outperforms the 3-class DNN model in five out of the seven metrics, yielding a mean accuracy of 0.55 across 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 official performance metric, obtaining a slightly lower murmur 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*. On the murmur detection task we obtained the scores of 0.467 (ranked 34th out of 40) and 0.374 (ranked 40th out of 40) on the hidden validation and test sets, respectively (Table 1). On the clinical outcome identification task, we scored 11032 (ranked 25th out of 39) and 18754 (ranked 39th out of 39) on the hidden validation and testing sets, respectively (Table 2). For both tasks, the results on the training set are reported based on patient-independent 10-fold cross-validation.

Training	Validation	Test	Ranking
0.56 ± 0.05	0.47	0.37	40/40

Table 1. Weighted accuracy metric scores for our final selected entry for the murmur detection task.

Training	Validation	Test	Ranking
6001 ± 1005	11032	18754	39/39

Table 2. Cost metric scores for our final selected entry for the clinical outcome identification task.

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

Table 3. Overall performance. Results presented are mean ± standard derivation for 10 folds.

4. Discussion and conclusions

With the release of a new HS 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 while the one-step approach achieved a better murmur score, it performed poorly on unknown detection. The proposed uncertainty-aware tandem learning, however, performed significantly better in unknown detection, and demonstrated a more balanced performance between murmur and normal detection.

Generally poor unknown sensitivity may have been caused by many “known” samples getting mislabelled as unknown; and, despite high specificity, the errors might have been introduced when combining results per patient. Therefore, future work could focus on exploring what reduces the model certainty on “known” samples, as well as alternative methods for combining results per patient.

In the present study, we prioritised identification of unknown samples due to its high value in practical applications. Early detection of samples with low model confidence could allow to prompt a user to repeat the recording in a less noisy environment or changing the stethoscope position, leading to a more accurate diagnosis.

The official challenge scores, however, place more weight on murmur detection. As a result, although our approach improved the model performance on the unknown class, the official performance attained by our approach was relatively poor. In addition, we observed notable performance differences between the validation and test sets for our method as well as others’. This might indicate a distribution mismatch between these two sets. In contrast, we performed cross-validation which might give a better insight on how a model would generalise.

While the main focus of this challenge was a three-class classification, the dataset’s metadata contains 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, et al. The PASCAL classifying heart sounds challenge 2011 (CHSC2011) results. http://www.peterjbentley.com/heartchallenge/index.html.

[2] Clifford GD, Liu C, Moody B, et al. 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] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, Jorge A, Mattos S, Hatem T, Tavares T, Elola A, Rad AB, Sameni R, Clifford GD, Coimbra MT. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[4] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, Perez-Alday EA, Sadr N, Sharma A, Mattos S, Coimbra MT, Sameni R, Rad AB, Clifford GD. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;.

[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 Proc. EMBC. 2021; 669–672.

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

Address for correspondence:

Erika Bondareva: eb729@cam.ac.uk
15 JJ Thomson Ave, Cambridge, UK, CB3 0FD