Exploring Composite Dataset Biases for Heart Sound Classification

Davoud Shariat Panah[1], Andrew Hines[2], and Susan Mckeever[1]

[1] School of Computing, Technological University Dublin, Ireland
[2] School of Computer Science, University College Dublin, Ireland
d19127274@mytudublin.ie, andrew.hines@ucd.ie, susan.mckeever@tudublin.ie

Abstract. In the last few years, the automatic classification of heart sounds has been widely studied as a screening method for heart disease. Some of these studies have achieved high accuracies in heart abnormality prediction. However, for such models to assist clinicians in the detection of heart abnormalities, it is of critical importance that they are generalisable, working on unseen real-world data. Despite the importance of generalisability, the presence of bias in the leading heart sound datasets used in these studies has remained unexplored. In this paper, we explore the presence of potential bias in heart sound datasets. Using a small set of spectral features for heart sound representation, we demonstrate experimentally that it is possible to detect sub-datasets of PhysioNet, the leading dataset of the field, with 98% accuracy. We also show that sensors which have been used to capture recordings of each dataset are likely the main cause of the bias in these datasets. Lack of awareness of this bias works against generalised models for heart sound diagnostics. Our findings call for further research on the bias issue in heart sound datasets and its impact on the generalisability of heart abnormality prediction models.

Keywords: Bias, PhysioNet Dataset, Heart Sound, Machine Learning

1 Introduction

Cardiac auscultation is a cost-effective and non-invasive technique that has been used by physicians to diagnose heart disease for over 200 years [15]. Auscultation involves listening and interpreting the patient's heart beat, typically using a stethoscope. However, the accuracy of this diagnostic method is influenced by different factors such as the auscultation skills of the clinicians and the capacity of the human auditory system to detect low-frequency sounds [20]. In recent times, the development of heart sound classification models for automatic detection of heart abnormalities has been an active area of research [13, 11, 4, 23, 16]. Given that the ultimate goal of such systems is to assist clinicians with their decision making, the generalisability of these models to unseen real-world data is of great importance.

One of the main causes of poor generalisation of predictive models is dataset bias [22, 21]. Unintended bias can be introduced into datasets at different stages

of the data collection and generation process. Consequently, the source of bias will differ across datasets, including historical, representation and measurement bias [18]. Supervised machine learning models are heavily influenced by the characteristics of the data they are trained on. Bias in the data may result in a suboptimal model which would be biased towards some particular features of the dataset [22]. While such a model might show a high level of accuracy on the dataset used for training and evaluation, it may not offer the same performance when deployed to production. In other words, the generalisability of such model can be affected by the fact that the training and the real-world data come from different distributions.

In addition to the diagnostically salient acoustic characteristics, heart sound recordings are susceptible to a variety of factors. These can be grouped as: human factors regarding the patient (e.g. age, resting/moving state, fitness levels); context and environmental factors (e.g. room noise, stethoscope placement); and system factors (e.g. stethoscope specifications such as acoustic coupling, digital sampling rate, acoustic dynamic intensity range). Recent work has also shown differences in parameters of the transmitted sound exist between digital stethoscopes [12]. Despite the fact that any of these factors can be a potential source of bias in heart sound datasets, the impact of bias on datasets used for data-driven classification models has not been explored. At the same time, the heavy reliance on a small set of datasets in this area of research also stresses the importance of potential bias in these datasets. Currently, there are few publicly available heart sound datasets which have been employed to build heart abnormality prediction models. One of these which has been widely used as a gold standard dataset by researchers since 2016 is the PhysioNet heart sound dataset [9]. We explore the presence of potential bias in the PhysioNet heart sound dataset and its impact on the generalisability of heart disease prediction models. Our key finding is that bias is present in the PhysioNet meta-dataset, and that the sound capturing sensor (the digital stethescope) used is likely the principal contributor to this bias.

Our paper is structured as follows: In section 2, we provide a brief overview of the heart sound classification problem. We also give an overview of the available datasets and specifications of sensors used to capture heart sounds. In section 3, we provide the details of the experimental methodology, including the chosen datasets, pre-processing, feature representations, classification model, and metrics. In section 4, we give a detailed analysis of the results. In section 5, the results are discussed. Conclusions and future directions are presented in section 6.

2 Background and Related Work

Heart sounds are a product of vibrations in heart muscles. These vibrations are in turn the result of blood flow with the opening and closure of heart valves. A normal heartbeat cycle is composed of two separate sounds, called first heart sound (S1) and second heart sound (S2). In some cases, a third and fourth sound might also be present, which can be a sign of heart abnormality. In addition

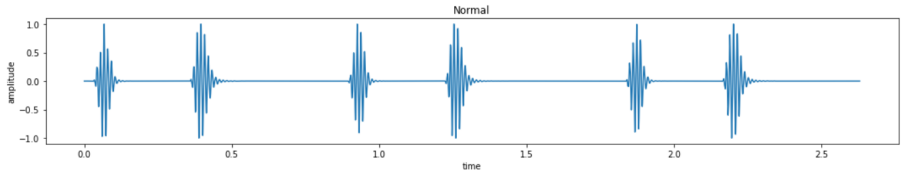

Fig. 1: Phonocardiogram of a normal heart sound.

to these sounds, heart valve defects can also produce a whooshing or swishing sound, which is called murmur. A phonocardiogram is a visual representation of a heart sound showing heart sound amplitudes over a period of time. Figure 1 shows the phonocardiogram of a normal heart sound.

Physicians use a device called stethoscope to monitor the heart sounds. By examining the timing, duration, intensity, and pitch of the heart sounds, they can differentiate normal and abnormal sounds [10]. Acoustic stethoscopes have been used by clinicians for over 200 years. However, in recent years, they have evolved into digital auscultation devices with multiple functionalities. 3M Littmann [1], Eko Core [6], and Jabes [7] are just some examples of available electronic stethoscopes in the market. Specifications of these three stethoscopes have been provided in Table 1. These devices allow their users to significantly amplify heart sounds. They also eliminate ambient noises by filtering out unwanted frequency ranges or through active noise cancellation. Littman and Jabes stethoscopes also offer functionality to enable users to switch between different frequency modes which have been tailored to heart and lung sounds frequency ranges. Such features enable digital stethoscopes to offer a higher level of sound quality than their acoustic counterparts, which in turn will assist practitioners to make a more accurate diagnosis [19]. Referring to Table 1, we think that characteristics such as different frequency ranges, digital sampling rates, and frequency modes can be potential causes of bias in sounds recorded by such sensors.

Table 1: Specifications of three digital stethoscopes available in the market.

Stethoscope	**Frequency Range (Hz)**	**Sample Rate (Hz)**	**Amplification**	**Noise Reduction**	**Frequency Modes**
3M Littmann 3200	20 – 2000	8000	Up to 24x	Yes	Yes
Eko Core	20 – 2000	4000	Up to 40x	Yes - Active noise cancellation	No
Jabes	20 – 1000	8000	Up to 20x	Yes	Yes

Currently, a few heart sound datasets are available to researchers, including but not limited to PASCAL [2] and PhysioNet datasets. PhysioNet is by far the most extensive heart sound meta-dataset, comprising six smaller databases with

different number of recordings. These six databases were contributed by different research groups to the PhysioNet Computing in Cardiology 2016 challenge. The heart sounds available in each database have been recorded using digital stethoscopes/microphones in clinical as well as non-clinical environments. PhysioNet meta-dataset contains 3240 recordings, out of which 665 samples belong to normal subjects and 2575 samples to abnormal ones. Since its release in 2016, many researchers, including [13, 11, 4, 23] have used PhysioNet as a benchmark dataset to validate their proposed algorithms.

Creating a heart sound classification system generally involves four steps: data acquisition and pre-processing, segmentation, feature extraction, and heart sound classification [5]. It must be noted that one or some of these steps might not be present in particular cases such as deep learning models. Potes et al. [13] applied the Springer segmentation algorithm [17] to segment heart sounds and then used time and frequency domain features to train an ensemble model which combines an AdaBoost classifier with convolutional neural network (CNN). Their method achieved a mean accuracy of 86% on the PhysioNet/CinC 2016 challenge test set and won the challenge. The goal of this challenge was to classify heart sound recordings into either normal or abnormal categories. In [11], Noman et al. extracted Mel-Frequency Cepstral Coefficients (MFCCs) from short segment heart sound signals. They then built a deep learning architecture which combines a 1D-CNN that receives raw heart sound signals, and a 2D-CNN that takes MFCCs as input. Their proposed method achieved a mean accuracy of 88.14% on the PhysioNet dataset. The method proposed by Dominguez-morales et al. [4] split heart sound recordings into fixed-length segments and extracts frequency bands of each segment. Sonogram images were generated for each of the samples and then fed into a CNN model. This method achieved a mean accuracy of 94.16% on Physionet dataset. High reliance on the PhysioNet dataset as a benchmark dataset in recent work [13, 11, 4, 23] indicates that PhysioNet is currently the gold standard dataset in this field.

Although some of the studies mentioned above have achieved high accuracies on PhysioNet meta-dataset, the potential presence of bias in this dataset and its impact on the generalisability of the proposed models has been overlooked. The fact that the PhysioNet meta-dataset is imbalanced across its constituent databases increases the risk that the models built using this dataset will be biased towards the characteristics of one of its sub-databases. As a result, while the resultant models might achieve high accuracies on this particular dataset, such models may show lower performance when we use them in real-world scenarios to classify heart sounds. Previous studies were motivated to design models with a higher levels of heart sound classification accuracy. We explore the PhysioNet dataset from a different point of view. We investigate the presence of bias in PhysioNet, the leading dataset in the field, aiming to establish the presence and main cause of such bias.

Table 2: Details of the databases used for training and evaluation of the model. Adapted from [9].

Database	Subject type	# Rec	Age	Gender (F/M)%	Recording position	Sample rate (Hz)	Sensor frequency response	Sensor
Training-a	Normal Abnormal	117 292	Unknown	Unknown	Nine different positions	44100	20 Hz – 20 kHz	Meditron
Training-b	Normal Abnormal	386 104	Unknown	38/62	Tricuspid Area	4000	20 Hz – 1 kHz	3M Littmann E4000
Training-f	Normal Abnormal	80 34	56 ± 16	62/38	Apex	8000	20 Hz – 1 kHz	Jabes

3 Experimental Methodology

Our first objective is to find out if there is any bias resulting from PhysioNet's construction through combining sub-datasets of heart sounds that were sourced from a variety of independent research studies. To do so, we train a classification model using three different sub-datasets of PhysioNet dataset and see if we can distinguish the recordings of each sub-dataset with an accuracy higher than random guess.

PhysioNet sub-datasets have some differences across multiple attributes such as age distribution of subjects, auscultation positions and sensors used to capture the sounds. As a result, any of these attributes might be a potential cause of bias in this meta-dataset. In this regard, our second objective is to find out which attributes play a more significant role in introducing bias in the PhyioNet dataset.

This section presents the datasets, pre-processing, feature representations, classification model and the evaluation metrics computed. The experiments were implemented in Python 3.8 using Librosa 0.8.0 library for feature extraction and Scikit-learn 0.23.2 for the classification models.

3.1 Datasets

In order to train and evaluate the classification model, we use three out of six databases which are available in PhysioNet heart sound meta-dataset. Table 2 shows the details of the selected databases.

These databases include both normal and abnormal heart sounds and been recorded using three different electronic stethoscopes (sensors). We excluded Training-c and Training-d databases because the number of normal samples in these databases are small. As the sensors which were used to capture heart sounds are different across the normal and abnormal classes of Training-e database, we also do not use this database. It must be noted that the sampling rate of the recordings available in PhysioNet dataset is 2000 Hz.

We label the recordings of each of these databases according to their sensors. Then we choose 80 recordings from the *normal* class of each of the three datasets. Out of each of these sets of 80 samples, 50 recordings will be used for training, and 30 recordings will be used for testing the classification model. Therefore, we will have a training set containing 150 *normal* samples across three different databases, and a test set of 90 *normal* samples. We also create a separate test set which includes 90 *abnormal* recordings from the same databases. In Table 3, the details of the training and test sets which are used to train and evaluate the classification model have been provided. After creating the training and test sets, pre-processing, feature extraction, and classification steps will be carried out as described in section 3.2, 3.3, and 3.4, respectively.

Table 3: Distributions of samples selected from each dataset in training and test sets. The training set contains only *normal* recording. The classes include Jabes, Littmann, and Meditron.

Sensor(class)	**Training set**	**Normal test set**	**Abnormal test set**
Jabes	50	30	30
Littmann	50	30	30
Meditron	50	30	30
Total	150	90	90

3.2 Pre-processing

Given that each heart sound recording has a different duration, to make the length of samples consistent, we use only the first five seconds of each sample. Five seconds is long enough to capture several cardiac cycles. Also, given that recordings of each database have a different range of amplitudes, we normalise all samples to have an amplitude between -1 and 1 with the following equation:

$$S\acute{}(t) = 2\frac{S(t) - min[S(t)]}{max\,[S(t)] - min\,[S(t)]} - 1 \tag{1}$$

where $S(t)$ and $S\acute{}(t)$ are the original and normalised signals, respectively.

3.3 Feature Extraction

Numerous time-domain, frequency-domain and time-frequency features have been employed in the area of heart sound classification [5]. Unlike previous work, in this paper we are not aiming to design a heart abnormality prediction model – our main goal is to determine whether combining training data from a variety of sources introduces bias into the dataset. As heart sounds can be classified by listening to them with a stethoscope, basic acoustic features that capture features salient to human perception are applied. Given that heart sounds are fundamentally periodic beats, both temporal and spectral features are captured. In this regard, after the pre-processing step, we extract four different spectral features from each sample, including spectral centroid, spectral roll-off, spectral

bandwidth and spectral contrast. To reduce the dimensionality of feature vectors, for all features, we calculate the average value of the feature across the frames of the sample, as in [23].

(a) *Spectral centroid* is a measure that shows where the centre of mass of the spectrum is located. It represents the brightness of the sound signal [14] and is calculated as follows (as used in [23]):

$$S_c = \frac{\sum_n \; x\left(n\right) f\left(n\right)}{x\left(n\right)} \tag{2}$$

where $x\left(n\right)$ represents the spectral magnitude of frequency bin n, and $f\left(n\right)$ is the centre frequency of that bin.

(b) *Spectral bandwidth* is the order-p statistic of the signal spectrum and distinguishes high bandwidth sounds from low bandwidth sounds. It is calculated using the following equation (as used in [14]):

$$S_b = \left(\sum_n x\left(n\right)\left(f\left(n\right) - S_c\right)^p\right)^{\frac{1}{p}} \tag{3}$$

where $x\left(n\right)$ is the spectral magnitude at frequency bin n, $f\left(n\right)$ is the centre frequency of that bin, and S_c represents the spectral centroid. It must be noted that in default Librosa implementation of this feature, p is equal to 2.

(c) *Spectral roll-off* point is a frequency so that 85% of spectral energy lies below that frequency. This feature is calculated using the following equation (as used in [23]):

$$\sum_{n=1}^{f} x\left(n\right) = 0.85\left(\sum_{n=1}^{N} x\left(n\right)\right) \tag{4}$$

where f is the roll-off frequency, $x\left(n\right)$ is the spectral magnitude at frequency bin n, and N is the total number of frequency bins.

(d) *Spectral contrast* is defined as the difference between spectral peaks and valleys measured in sub-bands by octave-scale filters. For more information, please refer to [8].

3.4 Classification Model

After the feature extraction step, we train a linear Support Vector Machine (SVM) classifier [3]. We selected SVM classifier based on the data volumes available, and achieved similar results with other classifiers such as KNN and random forest. We perform a grid search with 4-fold cross-validation on the training set (as described in Table 3) to optimise the c value for the SVM. After training, the model is evaluated using the test sets described in Table 3.

3.5 Metrics

We use two different metrics to evaluate the classification model. The first one is recall, which is also called sensitivity or true positive rate. It shows the fraction of positive examples which have been classified correctly and is calculated using the following equation:

$$Recall \ = \frac{True \ positive}{True \ positive \ + \ False \ negative}. \tag{5}$$

As summarised in Table 3, the prepared training and test sets are balanced across classes. Therefore, we also use accuracy metric to evaluate our model at dataset level. This metric is defined as the ratio of the number of correct predictions to the total number of examples and is calculated as follows:

$$Accuracy \ = \ \frac{True \ positive \ + \ True \ negative}{All \ examples}. \tag{6}$$

4 Results

4.1 Investigation of the presence of dataset bias

Figure 2 illustrates the distributions of the extracted features from the training set across three different datasets. As shown in Figure 2, the median value of the majority of features is distinct across different datasets. Also, in the case of Jabes dataset, we can observe that the distributions of spectral centroid and spectral roll-off features are entirely distinct from the feature distributions of the other two datasets. The observations can be a sign of potential bias across datasets.

To find out if selected datasets are biased or not, we evaluate the SVM model which was trained using *normal* recordings of three different datasets on the *normal* test set (as described in Table 3). Given that normal recordings must be consistent across different datasets, if dataset bias was not present, we could not expect to see an accuracy higher than chance. Figure 3 (left confusion matrix) depicts the evaluation results. According to Figure 3, we can see when we evaluate the model on the test set with *normal* recordings, the recall for each of the classes is at least 97%. Given that this is a three-class classification problem, the random guess accuracy would be $\frac{1}{3} \approx 0.33$. We can see that the overall accuracy on the *normal* test set is 98% which is significantly higher than the chance. This observation clearly indicates the presence of bias in selected PhysioNet sub-datasets.

4.2 Exploring the cause of the dataset bias

As we mentioned earlier, in addition to sensors, PhysioNet sub-datasets are different across multiple attributes such as age distribution of the subjects and auscultation positions used to record the heartbeat sounds. To find out what is

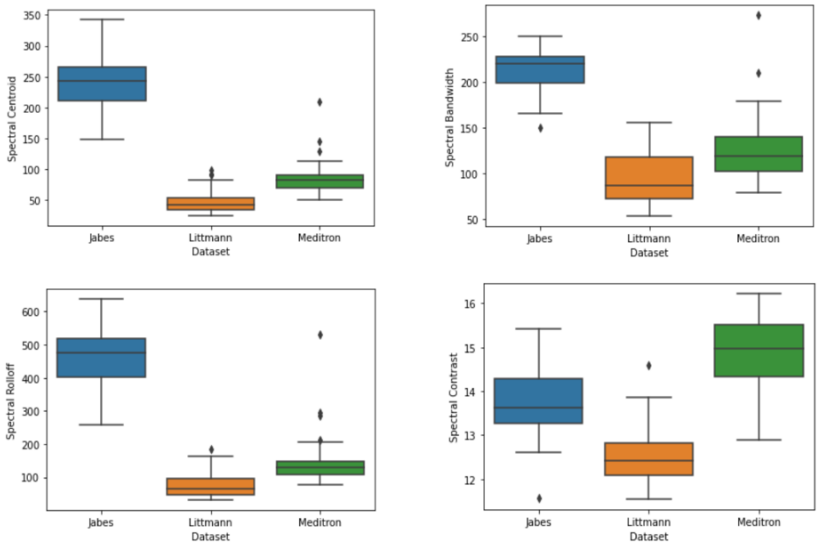

Fig. 2: Distributions of extracted features across three different datasets.

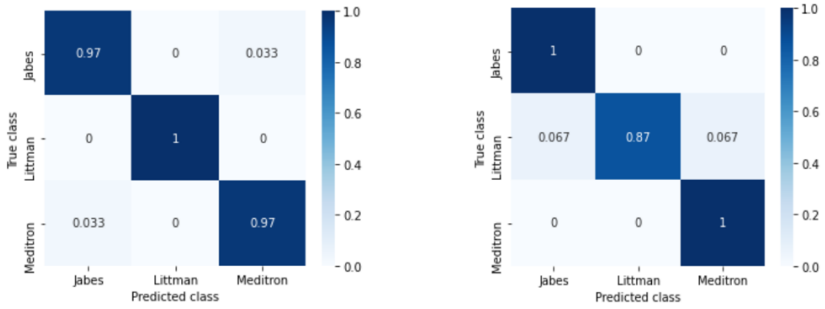

Fig. 3: Confusion matrices for the SVM model which has been trained using four spectral features. The left confusion matrix shows the performance of model tested on *normal* test set, and the right one shows the performance on the *abnormal* test set.

the main cause of bias in PhysioNet dataset, we perform the above experiment again, but this time instead of testing on the *normal* test set, we evaluate the model on the *abnormal* test set (as described in Table 3). In other words, we use the same SVM model which was trained on the *normal* heart sounds and evaluate it on a test set which contains various *abnormal* heart sounds from different subjects. This way, we can make sure that the recordings in training and test sets are considerably different in terms of content. If the model can still predict the datasets with an accuracy higher than chance, this will indicate that sensor is likely the main cause of bias in datasets and the role of the other

attributes in dataset bias is negligible compared to that of sensors. The reason is that sensor is the only attribute which is certainly consistent across the three datasets available in normal and abnormal test sets.

Figure 3 (right confusion matrix) shows the result of this experiment. We can see that the recall values for all three classes are still much higher than the random guess. Also, the overall accuracy is 96% which is significantly higher than the chance (33.3%). We observe that despite evaluating the model on a test set with very different content, the model still achieves a near to perfect accuracy. This observation suggests that the sensor is likely the main source of bias in selected heart sound databases. Given that other attributes like the age distribution of the subjects differ across *normal* and *abnormal* test sets, if they were the main source of bias, we could not except to see an accuracy much higher than chance.

5 Discussion

In section 4, we examined the presence of bias across three heart sound datasets available as part of the PhysioNet meta-dataset We showed that we could accurately classify sub-datasets of the PhysioNet meta-dataset and concluded that bias is certainly present in this meta-dataset. We also demonstrated that the role of attributes such as the age of the subjects and auscultation positions in introducing bias could not be as significant as the sensor, and sensor is likely the main cause of bias in PhysioNet dataset. It is important to note that, due to the existence of multiple attributes in each of the PhysioNet sub-datasets, we cannot assert that the other attributes do not play any role in introducing bias into this dataset. We do not have access to sufficient data to precisely examine the role of all attributes involved.

We carried out our experiments on three out of six subset datasets of PhysioNet meta-dataset. This meta-dataset has been assembled by pooling smaller datasets from different resources. As it was reported in Table 2, each of these datasets contains a different number of recordings. This means that when we use PhysioNet meta-dataset to train heart abnormality prediction models, we can expect a bias in our models towards the characteristics of databases with the highest number of samples. That is to say, employing PhysioNet dataset for training heart disease prediction models may not necessarily lead to models with better generalisability than that of models trained with smaller datasets as the models can be biased towards a proportion of the PhyioNet meta-dataset. It is worth noting that PhysioNet meta-dataset has been used as a gold standard dataset in the majority of studies in the area of heart sound classification since 2016. The main goal of such studies is to build prediction models which can be used as a tool for initial screening of heart disease. However, according to the results of our experiments, the generalisability of the models built using this meta-dataset to unseen data in real-world settings seems implausible. Indeed, any model which is built using this meta-dataset must be evaluated using real-world data to validate that it can reproduce the reported performance. In addition to

this, we must also consider bias as an important factor when we want to create a heart abnormality prediction system using the PhysioNet meta-dataset. As we mentioned in section 2, in the majority of cases, building a heart abnormality prediction model involves four different steps: pre-processing, segmentation, feature extraction, and classification. Our design decisions in each of these steps can determine the level of which the resultant model will be influenced by the dataset bias.

6 Conclusion and Future Work

In this paper, we investigated the presence of potential bias in PhysioNet heart sound meta-dataset. We chose three sub-datasets of this dataset and labelled the recordings of each one based on the sensor used to capture them. Then we built an SVM model using four spectral features. The model was able to detect recordings of each of the PhysioNet sub-datasets with an accuracy of 98%, which is way above chance. This indicates that bias is undoubtedly present in PhysioNet dataset, the gold standard dataset in the field. We also showed that sensors are likely the main cause of this bias. Our findings necessitate further investigations into the impact of this bias issue in the PhysioNet dataset on the generalisability of the heart sound classification models.

A comprehensive analysis of the different feature representations which are being used in the field of heart abnormality prediction in terms of their level of robustness to sensor bias can be a future direction. In addition to this, looking into the possibility of alleviating the bias through data preprocessing techniques can also be an interesting future work.

Acknowledgements This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224.

References

1. 3M: Littmann electronic stethoscope model 3200, https://www.littmann.com/3M/en_US/littmann-stethoscopes/, last accessed 2020/09/16.
2. Bentley, P., Nordehn, G., Coimbra, M., Mannor, S.: The pascal classifying heart sounds challenge (2011), http://www.peterjbentley.com/heartchallenge/index.html
3. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152 (1992)
4. Dominguez-Morales, J., Jimenez-Fernandez, A., Dominguez-Morales, M., Jimenez-Moreno, G.: Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. IEEE Trans. Biomed. Circuits Syst **12**, 24–34 (2018)

5. Dwivedi, A., Imtiaz, S., Rodriguez-Villegas, E.: Algorithms for automatic analysis and classification of heart sounds–a systematic review. IEEE Access **7**, 8316–8345 (2019)
6. Ekohealth: Core digital stethoscope - electronic stethoscopes — eko, https://shop.ekohealth.com/products/core-digital-stethoscope, last accessed 2020/09/16.
7. Jabes: Jabes electronic stethoscope, https://www.allheart.com/jabes-electronic-stethoscope/p/jsjabes3/, last accessed 2020/09/16.
8. Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H.: Lian-hong cai: Music type classification by spectral contrast feature. In: Proceedings. IEEE International Conference on Multimedia and Expo. p. 113–116. IEEE, Lausanne, Switzerland (2002)
9. Liu, C., Springer, D., Li, Q., Moody, B., Juan, R., Chorro, F., Castells, F., Roig, J., Silva, I., Johnson, A., Syed, Z., Schmidt, S., Papadaniil, C., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M., Sameni, R., Mark, R.: Clifford, g.d.: An open access database for the evaluation of heart sound algorithms. Physiol. Meas **37**, 2181–2213 (2016)
10. McGee, S.: Auscultation of the Heart: General Principles. Evidence-Based Physical Diagnosis. Elsevier Health Sciences, 4th edn. (2017)
11. Noman, F., Ting, C.M., Salleh, S.H., Ombao, H.: Short-segment heart sound classification using an ensemble of deep convolutional neural networks. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP. p. 1318–1322 (2019)
12. Nowak, L.J., Nowak, K.M.: Sound differences between electronic and acoustic stethoscopes. BioMedical Engineering OnLine **17**(1), 104 (2018)
13. Potes, C., Parvaneh, S., Rahman, A., Conroy, B.: Ensemble of feature based and deep learning-based classifiers for detection of abnormal heart sounds. In: 2016 Computing in Cardiology Conference September. vol. 14 (2016)
14. Sharma, G., Umapathy, K., Krishnan, S.: Trends in audio signal feature extraction methods. Applied Acoustics **158**, 107020 (2020)
15. Shaver: J.a.: Cardiac auscultation: A cost-effective diagnostic skill. Current Problems in Cardiology **20**, 447–530 (1995)
16. Son, G.Y., Kwon, S., et al.: Classification of heart sound signal using multiple features. Applied Sciences **8**(12), 2344 (2018)
17. Springer, D.B., Tarassenko, L., Clifford, G.D.: Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering **63**(4), 822–832 (2015)
18. Suresh, H., Guttag: J.v.: A framework for understanding unintended consequences of machine learning (2020), arXiv:1901.10002 [cs, stat].
19. Tavel: M.e.: Cardiac auscultation: a glorious past–and it does have a future! Circulation **113**, 1255–1259 (2006)
20. Tavel, M.E.: Cardiac auscultation: a glorious past—but does it have a future? Circulation **93**(6), 1250–1253 (1996)
21. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars: T.: A deeper look at dataset bias (2015), arXiv:1505.01257 [cs].
22. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
23. Yadav, A., Singh, A., Dutta, M.K., Travieso, C.M.: Machine learning-based classification of cardiac diseases from pcg recorded heart sounds. Neural Computing and Applications pp. 1–14 (2019)