

2023

Exploring the Impact of Noise and Degradations on Heart Sound Classification Models

Davoud Shariat Panah

Technological University Dublin, davoud.x.shariatpanah@mytudublin.ie


Andrew Hines

Technological University Dublin, andrew.hines@tudublin.ie

Susan McKeever

Technological University Dublin, susan.mckeever@tudublin.ie

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomart>

 Part of the [Computer Engineering Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Shariat Panah, Davoud; Hines, Andrew; and McKeever, Susan, "Exploring the Impact of Noise and Degradations on Heart Sound Classification Models" (2023). *Articles*. 186.
<https://arrow.tudublin.ie/scschcomart/186>

This Article is brought to you for free and open access by the School of Computer Science at ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, vera.kilshaw@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-Share Alike 4.0 International License](#).

Funder: Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.



Exploring the impact of noise and degradations on heart sound classification models

Davoud Shariat Panah^{a,*}, Andrew Hines^b, Susan McKeever^a

^a School of Computer Science, Technological University Dublin, Dublin, Ireland

^b School of Computer Science, University College Dublin, Dublin, Ireland

ARTICLE INFO

Dataset link: <https://arrow.tudublin.ie/datas/20/>

Keywords:

Phonocardiogram
Heart sound
Heart sound classification
Noise and degradation
Quality enhancement
Synthetic dataset

ABSTRACT

The development of data-driven heart sound classification models has been an active area of research in recent years. To develop such data-driven models in the first place, heart sound signals need to be captured using a signal acquisition device. However, it is almost impossible to capture noise-free heart sound signals due to the presence of internal and external noises in most situations. Such noises and degradations in heart sound signals can potentially reduce the accuracy of data-driven classification models. Although different techniques have been proposed in the literature to address the noise issue, how and to what extent different noise and degradations in heart sound signals impact the accuracy of data-driven classification models remains unexplored. To answer this question, we produced a synthetic heart sound dataset including normal and abnormal heart sounds contaminated with a variety of noise and degradations. We used this dataset to investigate the impact of noise and degradation in heart sound recordings on the performance of different classification models. The results show different noises and degradations affect the performance of heart sound classification models to a different extent; some are more problematic for classification models, and others are less destructive. Comparing the findings of this study with the results of a survey we previously carried out with a group of clinicians shows noise and degradations that are more detrimental to classification models are also more disruptive to accurate auscultation. The findings of this study can be leveraged to develop targeted heart sound quality enhancement approaches — which adapt the type and aggressiveness of quality enhancement based on the characteristics of noise and degradation in heart sound signals.

1. Introduction

Cardiovascular diseases are currently the leading cause of mortality worldwide, accounting for one-third of deaths globally [1]. Early diagnosis through pervasive approaches can help detect heart disease in patients at earlier stages and consequently improve the survival rate. Auscultation has been a cost-effective approach for pre-screening heart disease for over 200 years [2]. However, auscultation is a subjective practice and requires extensive training [3]. Therefore, automatic analysis of heart sounds has been presented as an alternative to auscultation for early pre-screening of heart abnormalities. In recent years, a variety of data-driven heart disease diagnostic systems have been developed that can distinguish normal from abnormal heart sounds [4–8]. Such data-driven heart disease detection models can be used as pre-screening tools in situations where access to trained medical professionals is limited, which is the case in many under-developed regions of the world. Unfortunately, over 75% of deaths due to cardiovascular diseases occur in such regions of the world [1]. Automatic analysis of heart sounds

can help with the early diagnosis of heart disease which, in turn, could reduce mortalities due to heart disease.

Heart sounds are generally captured using digital stethoscopes or mobile phones. Although such devices typically benefit from noise reduction and cancellation technologies [9–11], they can still capture a considerable amount of noise while recording heart sounds, especially in noisy environments. Due to the presence of internal physiological body noises and ambient artifacts in clinical and non-clinical settings, it is almost impossible to record noise-free heart sound signals in real-world scenarios.

It has been stated that noise and contaminations in heart sound recordings can reduce the performance of data-driven models [12–15]. Researchers have adopted different approaches to mitigate the negative impact of noise and degradations in captured signals on the performance of data-driven heart disease diagnostic systems. Heart sound quality enhancement is one of the most widely adopted approaches to reducing noise in captured signals. A whole host of enhancement techniques have been employed in the field, such as filtering [4,5]

* Corresponding author.

E-mail addresses: davoud.x.shariatpanah@mytudublin.ie (D. Shariat Panah), andrew.hines@ucd.ie (A. Hines), susan.mckeever@tudublin.ie (S. McKeever).

<https://doi.org/10.1016/j.bspc.2023.104932>

Received 23 November 2022; Received in revised form 16 March 2023; Accepted 5 April 2023

Available online 14 April 2023

1746-8094/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

and wavelet-based denoising [14–17]. Heart sound quality classification is another approach in which a data-driven model distinguishes low-quality heart sounds from good-quality ones [18–20]. Low-quality signals are discarded, while recordings with an acceptable quality are subsequently presented to heart disease diagnostic models.

While numerous methods have been proposed in the field to enhance or classify the quality of heart sounds, it has remained unexplored how and to what extent noise and degradation in heart sound signals can impact the overall accuracy of data-driven models. A deeper understanding of the impact of noise and degradations in heart sound recordings on the performance of the data-driven models allows us to adapt the heart sound capture process with the aim of minimizing the negative impact of such noise and degradations. Also, such an understanding enables us to adjust the quality enhancement of the captured heart sounds based on their noise content and develop targeted heart sound pre-processing pipelines. In this regard, this study aims to answer the following research question: how do noise and degradations in heart sound signals impact the overall accuracy of data-driven models?

To answer this research question, we produce a synthetic dataset containing normal and abnormal heart sounds with various noises and degradations. This dataset will then be employed to train and evaluate multiple heart sound classification models. This enables us to systematically investigate the impact of a variety of noises and degradations on the performance of heart sound classification models. Previously we investigated the impact of noise and degradations on heart sound signals' diagnosability by conducting a survey with a group of trained clinicians [21]. In this study, we will also observe the similarities between the impact of noise and degradations on the performance of classification models and the results of our previous study.

The remainder of this paper is structured as follows: Section 2 overviews the related work on the impact of noise and the application of synthetic datasets. Section 3 provides the details of the datasets and data-driven models employed in this study. In Section 4, the results are given. In Section 5, results are discussed. Conclusions and future directions are presented in Section 6.

2. Related work

2.1. Noise impact

Noise and degradations with internal or external sources can have a negative impact on auscultation. Shindler [22] has stated that high levels of environmental sounds, such as speech, can interfere significantly with auscultation. Coviello [23] has pointed out that noises due to muscular movements can interfere with heart sounds and make it harder for clinicians to perceive the salient characteristics of the heart sounds. In addition to ambient and movement noises, internal physiological noises can also be disruptive to auscultation. Ranganathan et al. [24] have emphasized that intense breathing noises can interfere with assessing heart sounds and reduce the accuracy of auscultation. In our previous work, we surveyed a group of thirteen clinicians to understand the characteristics of diagnosable heart sounds and the impact of noises and degradations on the diagnosability of heart sound recordings [21]. This survey included a subjective listening test with 20 heart sound recordings contaminated with different noises and degradations. Analyzing the results of this survey showed that, from the point of view of clinicians, noise and degradations in heart sound signals have a detrimental effect on the diagnosability of heart sounds.

Noise and degradation in heart sound recordings can also reduce the accuracy of data-driven classification models. Paul et al. [12] have indicated that internal or external noises can mask fundamental heart sounds and increase the false positives of the classifiers. According to Kumar et al. [13], noise in heart sound recordings can alter the morphological characteristics of the heart sounds and change the features salient to accurate diagnostics. Gradolewski et al. [14] have stated

that some heart sounds, such as late-systolic and pan-systolic murmurs, have similar characteristics to noise, and, as a result, applying denoising algorithms can decrease the misclassification of such signals by data-driven models. Jain et al. [15] have indicated that noise and degradations in heart sound signals can reduce the accuracy of the segmentation of heart sounds into heartbeat cycles, which in turn can lead to a sub-optimal heart sound classification model. Although it has been emphasized that noise and degradation in heart sound signals can potentially reduce the accuracy of classification models, we could not find any comprehensive study exploring the impact of different noises and degradations on classification models.

2.2. Synthetic data

Synthetic data has been widely employed to develop and evaluate data-driven models in different domains. Synthetic data can have different use cases. For example, it has been used in cases where access to large datasets was limited or real-world datasets lack diversity [25–27]. It has also been used in cases where the impact of different variations in input data on the performance of algorithms has been the subject of the study [28].

In this study, we produced a heart sound dataset which is a synthetic combination of real-world clean heart sounds with a variety of noises in different SNR levels. Such a synthetic dataset offers several advantages over publicly available datasets for our use case. First, by synthetically adding noise to heart sound signals, we can generate recordings contaminated with a large variety of noise types common in both clinical and non-clinical settings. Also, by using different SNR levels, we can control the intensity of noise contamination in each of the recordings, allowing us to generate samples with various noise levels, from roughly clean to very noisy. Such a controlled synthetic setting enables us to thoroughly investigate the impact of different noise variables, such as noise types, groupings, intensities, and durations, on the performance of data-driven models. To date, publicly available heart sound datasets have been mainly captured in controlled environments, and recordings of such datasets are not diverse enough in terms of noise types and intensities. Also, it is very difficult to accurately determine the amount and types of noise in heart sound recordings in real-world datasets, and as a result, it would not be possible to provide a detailed analysis of the impact of noise and degradations on the performance of data-driven models using such datasets.

3. Method

In this section, we provide the details of the methodology for this study. The overall approach can be summarized as follows:

- A synthetic heart sound dataset including normal and abnormal heart sounds contaminated with a large variety of noise and degradations is generated. To generate this dataset, clean heart sounds are mixed with noises of different types, durations and groupings in different SNR levels.
- The synthetic dataset is split into train and test sets.
- Multiple classification models are developed. To develop these models, different feature representations (log-spectrogram and mel-spectrogram) and two commonly used classifiers in heart classification (support vector machine and convolutional neural network) are employed.
- Support vector machine models are trained using the synthetic training set. Convolutional neural network models are pre-trained using a dataset called PhysioNet and then fine-tuned using the synthetic training set.
- After training the models, we use the synthetic test set to evaluate the classification models.

Table 1

Details of the clean heart sound recordings used to generate the synthetic dataset.

Recording #	Type	Duration (s)
1	Normal	12.0
2	Normal	10.1
3	Normal	13.8
4	Normal	10.0
5	Normal	3.0
6	Normal	12.8
7	Normal	10.2
8	Normal	15.0
9	Abnormal - Aortic regurgitation	12.0
10	Abnormal - Aortic stenosis	10.9
11	Abnormal - Mitral regurgitation	12.0
12	Abnormal - Mitral stenosis	11.2
13	Abnormal - Mitral valve prolapse	11.5
14	Abnormal - Mitral valve prolapse	2.5
15	Abnormal - S3	10.1
16	Abnormal - S4	10.0

Table 2

Details of the noise types, their groupings and durations that have been mixed with clean heart sound recordings to generate the synthetic dataset.

Noise type	Noise grouping	Noise duration
White	Color	Long
Pink	Color	Long
Red	Color	Long
Sensor movement	Movement	Short
Body movement	Movement	Short
Deep breathing	Internal	Long
Fast breathing	Internal	Long
Coughing	Internal	Short
Digestive sound	Internal	Short
Talking	Ambient	Long
Door open/close	Ambient	Short
Phone ringing	Ambient	Long
Music	Ambient	Long
Water flow	Ambient	Long
TV	Ambient	Long
Dishwasher	Ambient	Long
Washing machine	Ambient	Long
Kettle	Ambient	Long
Vacuum cleaner	Ambient	Long
Dog barking	Ambient	Short
Bird singing	Ambient	Long

- To investigate the impact of noise and degradations in heart sound signals on the performance of classification models, we report the overall accuracies of the models across heart sounds contaminated with different noise types, durations, groupings and SNR levels.

In Section 3.1, we describe the process of generating the synthetic heart sound dataset. Also, we provide the details of another heart sound dataset called PhysioNet used in our experiments. Afterwards, in Section 3.2, we explain the stages of developing heart sound classification models, including pre-processing, feature extraction and classification.

3.1. Datasets

3.1.1. Synthetic dataset

This section provides the details of the synthetic heart sound dataset used in subsequent experiments and the stages of generating this dataset. Table 1 summarizes the specifications of the clean heart sounds, including their types and durations used to produce the synthetic dataset.

As shown in Table 1, we collected multiple clean normal, and abnormal heart sounds from different resources such as publicly available datasets and YouTube. There are sixteen clean heart sounds, consisting of eight normal and eight abnormal recordings. Abnormal recordings include more common murmurs and extra heart sounds. We chose

abnormalities that are more prevalent in publicly available heart sound datasets. As shown in Table 1, out of these sixteen recordings, fourteen signals are over 10 s long, and the duration of the other two signals is 2.5–3.0 s. Given that we aim to explore the impact of heart sound duration on the accuracy of the classification models, we included both short- and long-duration signals. Short-duration recordings are long enough to include at least two heartbeat cycles, but, at the same time, they are significantly shorter than the majority of the signals. We assessed the quality of these heart sound recordings through listening and visual inspection of the waveforms to ensure they are noise-free or contain a very low noise level.

Twenty-one different noise types were mixed with each of the base clean heart sound recordings. To have a comprehensive set of noise types, we chose the noise types that are common in clinical and non-clinical environments, such as home-places. Table 2 summarizes the details of the noise types, their groupings and durations. Noise types are categorized into four groups based on their source: color, movement, internal, and ambient. This categorization of noise types is similar to the classification provided by Gradolewski et al. [29]. Color noises were generated through simulation, while internal and ambient noises were collected from different publicly available datasets. Movement noises were captured using a mobile phone from the body surface. Regarding the ambient noises, noise types prevalent in clinical and non-clinical environments such as homeplaces were used. Including the noise types that are specific to non-clinical environments (e.g., home appliances noise) allows us to simulate situations where patients capture their heart sounds using consumer devices like mobile phones at home. Each clean heart sound was additively mixed with each noise contamination in ten different SNR levels: -10, -5, 0, 5, 10, 15, 20, 25, 30, and 40. In order to produce noisy heart sound signals with desired SNR levels, we changed the noise variance and additively mixed that with clean heart sound recordings. As shown in Table 2, these noise types are also categorized in terms of length into short- and long-duration noises. Long-duration noises were longer than the base clean heart sound signals, and as a result, they covered the whole or most parts of the generated recordings. In the case of short-duration noises, they were randomly distributed in time. In other words, we sampled a uniformly random number based on the duration of the clean heart sound recording and used that number as the starting point to add short-duration noise to the clean signal. Fig. 1(a) illustrates the phonocardiogram of a clean normal heart sound, and Fig. 1(b) shows the phonocardiogram of the same heart sound contaminated with dishwasher noise where SNR is equal to 10.

Using the process described above, 3360 synthetic heart sound recordings were generated. The sampling frequency of the recordings is 2000 Hz. Heart sounds samples and labels can be accessed online.¹ The synthetic dataset includes 210 noisy permutations for each clean heart sound recording. The specifications of the synthetic train and test sets are as follows:

- Half of the samples in the synthetic dataset (1680 recordings) are placed in the train set, and the other half in the test set.
- Train and test sets contain noisy permutations of different clean heart sound recordings: noisy permutations of heart sound numbers 1, 3, 6, 8, 9, 11, 12 and 15 are placed in the train set, while noisy permutations of heart sound numbers 2, 4, 5, 7, 10, 13, 14 and 16 are placed in the test set. The details of these base clean heart sound recordings have been provided in Table 1.
- Train and test sets are balanced across the two classes (normal and abnormal).
- The train set contains only long-duration recordings, while the test set contains both short- and long-duration signals.
- Noise types are the same across the train and test sets and include all twenty-one noise types, as summarized in Table 2.
- SNR levels are the same across the train and test sets and include ten levels: -10, -5, 0, 5, 10, 15, 20, 25, 30, and 40.

¹ <https://arrow.tudublin.ie/datas/20/>.

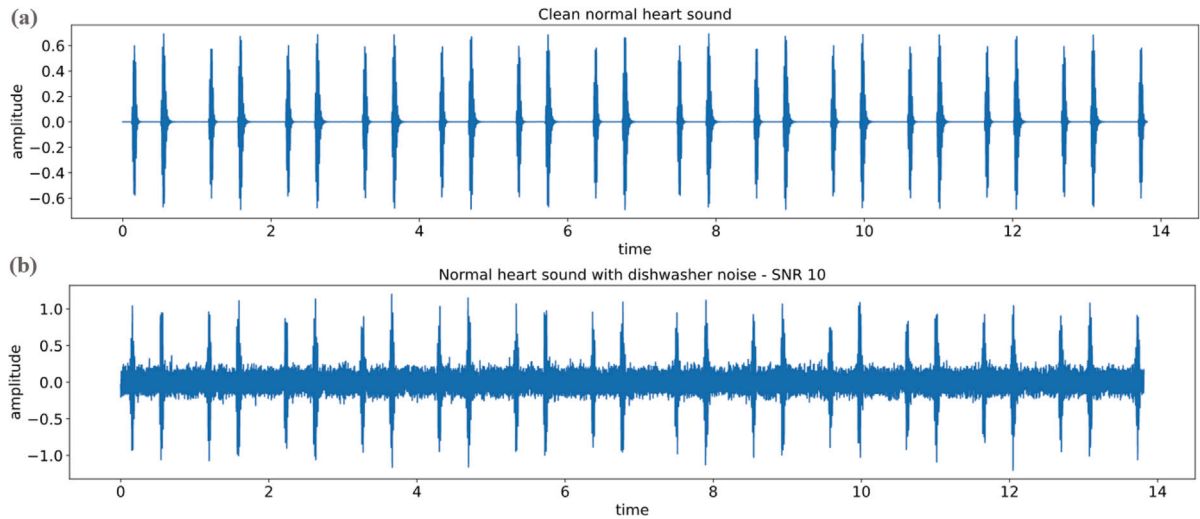


Fig. 1. (a) Phonocardiogram of a clean normal heart sound, (b) Phonocardiogram of the same heart sound contaminated with dishwasher noise.

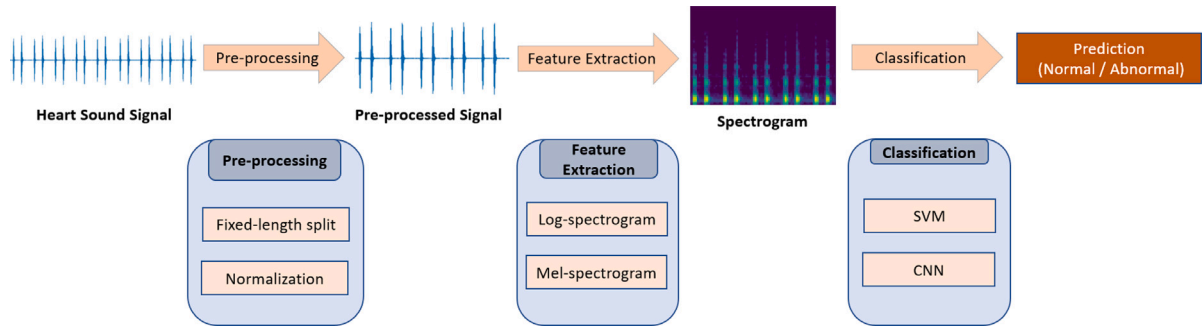


Fig. 2. Heart sound classification pipeline.

3.1.2. PhysioNet dataset

The PhysioNet heart sound dataset [30] was published as part of the PhysioNet/Computing in Cardiology 2016 challenge. This dataset comprises six smaller subset datasets that different research groups collected across the world in controlled or uncontrolled environments. PhysioNet dataset contains 3240 heart sound recordings, out of which 2575 samples were captured from healthy subjects while 665 samples were collected from pathologic subjects. Some of the recordings were labeled as unsure, which means that they were too noisy to be categorized as normal or abnormal. For this study, we excluded these low-quality recordings from the dataset. It should be noted that this dataset does not provide any information regarding the noise content (e.g., noise type and intensity) of the recordings. In the last few years, PhysioNet dataset has been widely employed as the largest publicly available heart sound dataset to develop data-driven heart sound classification models. This dataset is used in our experiments for pre-training deep learning models.

3.2. Data-driven models

As shown in Fig. 2, developing data-driven models for heart sound classification includes three steps: pre-processing heart sound recordings, extracting features from heart sounds, and training classification models. This section provides the details of these three stages.

3.2.1. Pre-processing

In the pre-processing stage, long-duration recordings are split into 5- or 10-s segments. 10-s and 5-s recordings are later used to train and test support vector machine (SVM) and convolutional neural network (CNN) models, respectively. Given that deep learning models generally

need a large number of samples for training, splitting the recordings into 5-s segments increases the number of available samples for training the deep learning models. The synthetic dataset also includes short-duration recordings which are only used for testing the models. These short-duration recordings are zero-padded before being fed into CNN models to ensure they are of the same length (5 s). Then, amplitude normalization is performed to minimize the variations in amplitudes across the signals, using the following equation (as in Ref. [16]):

$$X_{norm}(t) = \frac{X(t)}{\text{Max}(|X|)} \quad (1)$$

In the above equation, $X(t)$ represents the value of the heart sound signal at the time t , and $\text{Max}(|X|)$ is the maximum of the absolute value of the heart sound signal.

3.2.2. Feature extraction

After pre-processing the recordings, Linear- and Mel-scaled Short-Time Fourier Transform (STFT) features are extracted from signals. STFT is the most widely used time–frequency feature representation for heart sound classification [31]. This feature representation is computed using the following equation [32]:

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-i\omega n} \quad (2)$$

Where $x[n]$ is the signal to be transformed and $\omega[n]$ is the window function (Hann window). After computing the STFT of the signals, spectrogram representations were computed using the following equation:

$$S(m, \omega) = |X(m, \omega)|^2 \quad (3)$$

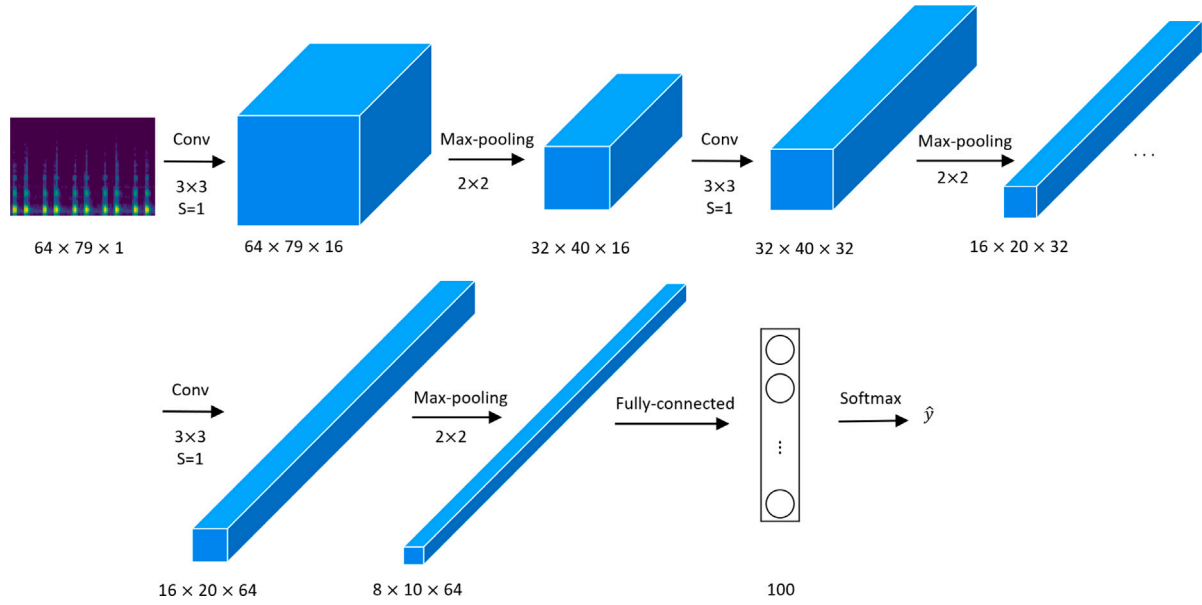


Fig. 3. Architecture of the CNN model with Mel-spectrogram as input (Mel-CNN model).

The spectrogram gives the power of the signal for each time and frequency pair. We used Log-spectrograms as well as Mel-spectrograms as our two feature representations. Mel-spectrogram is computed by converting the linear frequency scale to the Mel scale using the following formula [33]:

$$\text{Mel}(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (4)$$

We used Librosa library [34] to extract the above features. Window and hop lengths were fixed at 256 and 128, respectively. As for Mel-spectrograms, 64 Mel bands were used. In order to reduce the computational cost of training the support vector machine models, the average values of the features across the time axis were computed (as in Ref. [35]).

3.2.3. Classification

After extracting the feature representations, they are used to train the classifiers. Two different classifiers are employed in this study: Support Vector Machine (SVM) and Convolutional Neural Network (CNN). Both classifiers have been frequently used in the field to develop heart sound classification models. Given that we use Log-spectrogram and Mel-spectrogram as input features for these classifiers, four different models are developed: (a) Log-SVM, (b) Mel-SVM, (c) Log-CNN, and (d) Mel-CNN.

To implement the SVM models, we used the default parameters as provided in the Scikit-Learn library [36]. The synthetic dataset was used to train and evaluate the SVM models.

CNN models were implemented using TensorFlow 2.8 deep learning library. Fig. 3 shows the architecture of the CNN model with the Mel-spectrogram as input. This model consists of three convolutional layers. The first, second and third convolutional layers have 16, 32 and 64 kernels, respectively. A kernel size of (3, 3) was used for all three convolutional layers. Also, the stride was fixed at 1, and the ReLu function was used as the activation function. Each convolutional layer is followed by a max-pooling layer with a pool size of (2, 2). To reduce overfitting of the models, a dropout layer with a rate of 0.5 was used after each max-pooling layer. After convolutional and max-pooling layers, a fully connected layer with 100 neurons and the ReLu activation function was used. After this fully connected layer, another dropout layer with a rate of 0.5 is placed. The final layer of this architecture is Softmax which outputs the probability distributions of the potential outcomes (normal or abnormal).

Table 3

Performance of the classification models on the synthetic test set.

Model	Recall (Normal) %	Recall (Abnormal) %	Accuracy %
Log-SVM	67.9	73.9	70.9
Mel-SVM	68.8	75.6	72.2
Log-CNN	89.4 ± 1.3	62.6 ± 1.8	76.0 ± 0.4
Mel-CNN	82.3 ± 0.5	83.1 ± 1.1	82.7 ± 0.4

To train the models, Adam optimization [37] with a learning rate of 0.001 and cross-entropy objective function were used. CNN models were first pre-trained on the PhysioNet dataset for 60 epochs. This way, we can ensure that the CNN models are pre-trained on a large variety of normal and abnormal heart sounds. Then, all layers except fully connected layers were frozen, and the models were fine-tuned using the synthetic training set for 10 epochs. The trained CNN models were evaluated on the synthetic test set. This process was repeated ten times, and average and standard deviation values were reported for each metric.

It is worth noting that we first tried to train CNN models from scratch without pre-training using only the synthetic dataset. However, we observed those models were overfitting the synthetic dataset to an extreme extent. As a result, we excluded them from this study.

3.2.4. Evaluation metrics

The performance of the models is measured using two different metrics. The first one is recall which is used to quantify the performance of models across each class and calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

Given that the synthetic test set used to evaluate models is balanced across normal and abnormal classes, we also use accuracy to measure the overall performance of the classification models. Overall accuracy is computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

In the above equations, TP , FP , TN , and FN are the number of true positive, false positive, true negative, and false negative samples in the results test set, respectively.

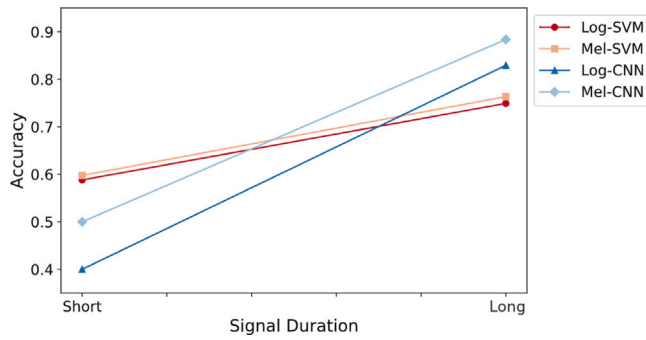


Fig. 4. Accuracy of the classification models across short- and long-duration heart sound recordings.

4. Results

4.1. Classification models' overall performance

Table 3 summarizes the results of evaluating classification models on the synthetic test set.

As shown in Table 3, CNN models outperform SVM models in terms of overall accuracy. The Mel-CNN model achieves the highest accuracy and recall for the abnormal class, while the Log-CNN model achieves the highest recall for the normal class.

4.2. Noise and degradation impact

In this section, we explore the impact of signal duration, noise type, noise grouping, noise duration and SNR on the accuracy of the classification models.

4.2.1. Signal duration

As mentioned in Section 3.1.1, the synthetic dataset contains short- (2.5–3.0 s) and long-duration (over 10 s) recordings. This dataset was split into training and test sets. The training set contains only long-duration signals, while the test set used to evaluate the classification models includes both short- and long-duration recordings. Fig. 4 depicts the impact of signal duration on the accuracy of the classification models.

As shown in Fig. 4, the overall accuracies of the models are considerably lower on short-duration heart sounds compared to long-duration ones. This drop in the performance of the models is more extreme for CNN models compared to SVM models. In other words, the accuracies of CNN models are over 80% on long-duration signals, while they fall below 50% when these models are evaluated using short-duration signals.

4.2.2. Noise type

As mentioned in Section 3.1.1, the synthetic dataset contains normal and abnormal heart sounds contaminated with twenty-one noise types. There are 80 samples in the test set for each noise type. Fig. 5 illustrates the overall accuracy of the models evaluated using heart sound signals contaminated with different noise types. In this plot, the noise types have been arranged based on the average accuracy of the classification models, from the noise type with the lowest (white noise) to the one with the highest (sensor movement noise) average accuracy.

As shown in Fig. 5, the classification models show different accuracies for heart sounds contaminated with different noise types. Also, we can observe that SVM models show larger fluctuations across different noise types than CNN models, indicating they are more sensitive to noise type than CNN models — We can see that for some noise types, like TV or dishwasher noise, SVM models offer the lowest accuracies (below 55%), while for some others, like deep breathing and phone ring noise, they achieve much higher accuracies (over 85%).

4.2.3. Noise grouping

As summarized in Table 2, noise types in synthetic heart sounds are categorized into four groups: color, movement, internal and ambient.

Fig. 6 compares the accuracy of the classification models across these noise groupings. As shown in Fig. 6, all classification models show their lowest performance on the recordings contaminated with color noises while performing best on the ones with movement noises. All models are sensitive to noise grouping, showing different accuracies across different noise groupings.

4.2.4. Noise duration

Noise contaminations used to generate the synthetic dataset can be categorized in terms of length into short- and long-duration noises (as specified in Table 2). Fig. 7 compares the accuracy of the classification models across recordings contaminated with short- and long-duration noises.

As shown in Fig. 7, all models are sensitive to noise duration and show lower accuracies when evaluated using signals contaminated with long-duration noises compared to short-duration noises.

4.2.5. SNR

As discussed in Section 3.1.1, in order to generate the synthetic dataset, clean heart sounds were mixed with noise contaminations with different SNR levels (from −10 to +40). For each SNR level, 168 recordings are available in the synthetic test set. Fig. 8 depicts the accuracy of the classification models across heart sound recordings with different SNR levels.

As shown in Fig. 8, Log-SVM and Mel-CNN models show a steady increase in accuracy from SNR −10 to 40. However, in the case of Mel-SVM and Log-CNN models, we observe an increase in accuracies from SNR −10 to 20, while for SNR levels higher than 20, the accuracies are roughly unchanged.

5. Discussion

5.1. Classification models' overall performance

In Section 4.1 of the results, we presented the overall performance of the classification models. The results show that the gap between the recalls of the normal and abnormal classes is larger for the Log-CNN model compared to the other models. This can indicate that the Log-CNN model is biased towards the normal class. This bias suggests that the Log-CNN model is overfitting the synthetic dataset. The synthetic dataset used to train the classification models was produced using a relatively small number of base clean heart sounds. At the same time, we should note that deep learning models such as CNNs generally need a large amount of training data, making them more prone to overfitting the synthetic dataset than SVM models. The reason why we only observe this overfitting in the case of the Log-CNN model may be that the higher dimensionality of the Log-spectrogram feature representation, at twice that of the Mel-spectrogram, makes the Log-CNN model more prone to overfitting than the Mel-CNN.

It is worth mentioning that drawing comparisons between the absolute accuracies of the classification models is not the purpose of this study. Instead, we aim to understand the pattern of accuracy change for each classification model in the face of different noise and degradation types.

5.2. Noise and degradation impact

In Section 4.2.1, we discussed the impact of heart sound signal duration on the performance of classification models. All models showed a lower performance when evaluated using short-duration recordings (2.5–3.0 s) than long-duration signals (5–10 s). This finding is in line with the results of our previous study with clinicians where the majority of the survey's respondents (92%) stated that they needed to listen to at

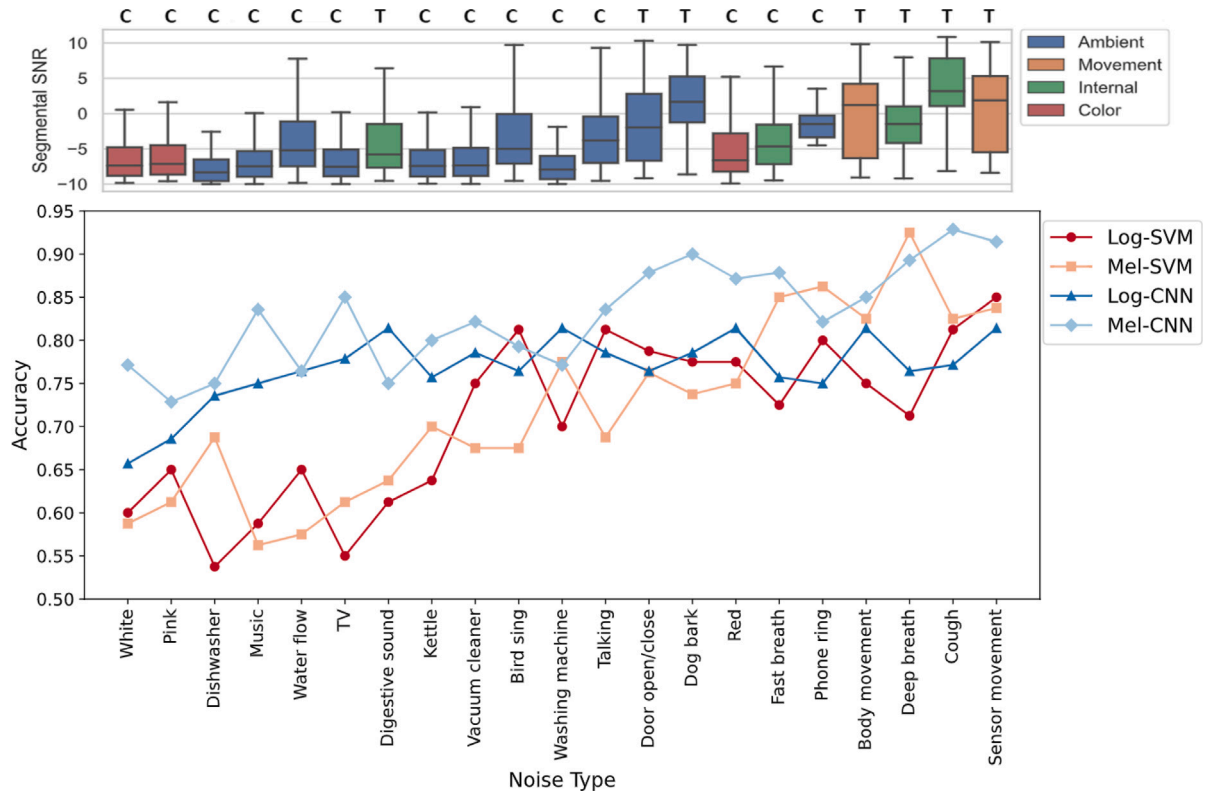


Fig. 5. Top: Range of the estimated segmental SNR across recordings contaminated with different noise types. Letters C and T stand for continuous (long-duration) and transient (short-duration), respectively. Bottom: Accuracy of the classification models across recordings contaminated with different noise types (lines are used for readability and ease of visualization but do not signify a relationship between categories on the x-axis.)

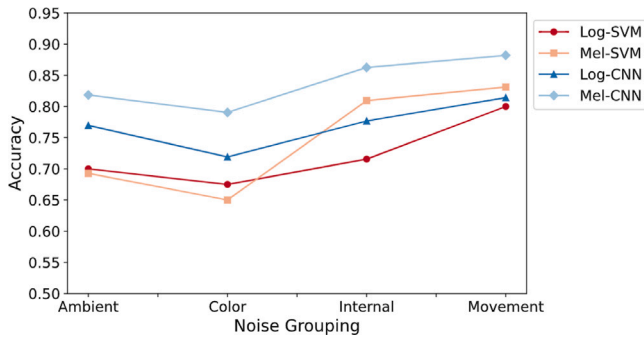


Fig. 6. Accuracy of the classification models across heart sounds contaminated with noises of four different groups. Lines are used for readability and ease of visualization but do not signify a relationship between categories on the x-axis.

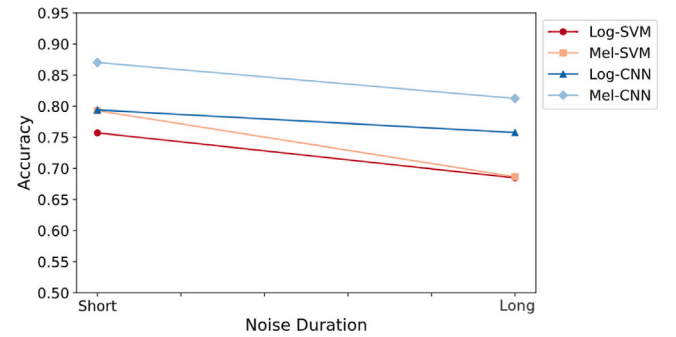


Fig. 7. Accuracy of the classification models across heart sounds contaminated with short- and long-duration noises.

least six heartbeat cycles before using the recordings for diagnosis [21]. As we mentioned in Section 3.2.1, we zero-padded the short-duration signals before using them for testing the CNN models to make sure the inputs to these models are of the same length. To rule out zero-padding as a contributing factor to the low accuracy of short-duration signals, we repeated the short-duration signals and used them to evaluate the CNN models. We observed similar results as in Fig. 4 which indicates that zero-padding does not reduce the accuracy of CNN models.

According to Chen et al. [38], an average heartbeat cycle is 0.8 s long, which means that the short-duration recordings in the synthetic dataset include around three heartbeat cycles on average. Short-duration heart sound signals, such as 1-s [6,38], or 3 s [39,40] recordings, have been used in the field to develop data-driven heart sound classification models. However, the results show that the classification models perform considerably worse on the short-duration recordings, which suggests that short-duration signals should be avoided

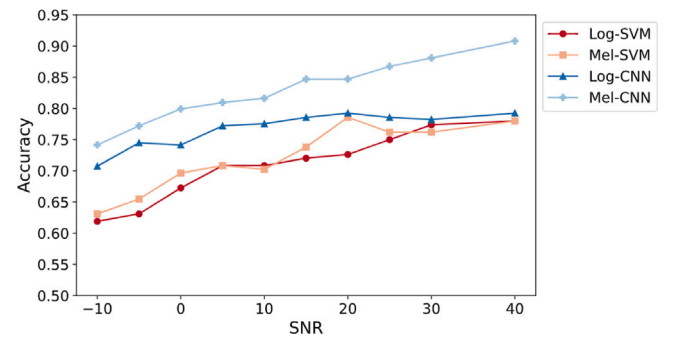


Fig. 8. Accuracy of the classification models across heart sounds with different SNR levels (from -10 to 40).

Table 4

Mean, standard deviation, minimum and maximum accuracies for each classification model across all recordings.

Model	Mean %	SD %	Min %	Max %
Log-SVM	70.9	9.5	53.8	85.0
Mel-SVM	72.2	10.5	56.3	92.5
Log-CNN	76.8	4.1	65.7	81.4
Mel-CNN	82.9	5.9	72.9	92.9

in situations where we can capture heart sound signals with longer durations.

In Section 4.2.2, we showed the performance of the classification models across heart sounds contaminated with varied noise types. Table 4 summarizes the mean, standard deviation, minimum and maximum accuracies for each classification model across all recordings in the test set. As shown in Table 4, for all noise types, the overall accuracy of the CNN models is over 65%, while in the case of SVM models, the accuracy goes below 57% for a few noise types. We can also see that the standard deviations of average accuracies are larger for SVM models compared to CNN models, which confirms the lower sensitivity of CNN models to noise types. However, what is common between these classification models is that they all react to the noise type, and they do not offer similar accuracies for all noise types. We can observe that there is only one short-duration noise (digestive sound) in the left half of Fig. 5, while in the right half, there are five short-duration noises (door open/close, dog bark, body movement, cough, sensor movement). Given that in this diagram, the noise types have been ordered based on the average accuracy of the classification models, this observation confirms that the continuous long-duration noises have a more detrimental effect on the accuracy of the models than transient short-duration artifacts. Also, we can see that most ambient and color noises are in the left half of the diagram while the majority of internal and movement noises are in the right half of the plot. This observation shows that ambient and color noises are more problematic than internal and movement noises for models.

Fig. 5 also illustrates the estimated segmental SNR values for recordings contaminated with different noise types. We can see that the majority of noise types on the left half of the diagram have smaller segmental SNR ranges with lower median values than the ones on the right half of the diagram. This observation confirms that signals contaminated with uniform noises are more challenging for classification models than the ones with transient noises. This can be due to the fact that continuous noises contaminate the entire signal. Transient noises, in contrast, contaminate a portion of the signal while the remaining parts can be of good quality. These results clearly show that both the duration and class of the noise are important factors that should be considered when analyzing the impact of noise on heart sound classification models. These results also suggest that different noise types should not be treated the same way by quality enhancement algorithms, as each noise type influences the classification models to a different extent.

In Section 4.2.3, we compared the performance of classification models on heart sounds contaminated with noises from different sources. We observed that color and ambient noises are more problematic for all models than movement and internal noises. This finding is also in agreement with the results of our previous survey with clinicians where respondents stated that ambient noises are more disruptive to accurate auscultation than internal or movement noises [21]. They also stated that internal and movement noises are roughly similar in terms of their negative impact on the diagnosability of heart sounds [21]. The observation that ambient and color noises have a more detrimental impact on the accuracy of the classification models than other noise types demonstrates that heart sounds contaminated by such noises should be prioritized over other noise sources for quality enhancement.

In Section 4.2.4, we explored the impact of noise duration on the accuracy of classification models. We saw that all models perform worse

on heart sounds contaminated with long-duration noises than the ones contaminated with short-duration noises. These results are also aligned with the survey results from our previous study. The survey results indicated that continuous long-duration noises are more disruptive than transient short-duration noises to accurate diagnostics [21]. This is intuitively correct given the longer duration of distortion of the heart signal and shows that long-duration noises must be prioritized over short-duration noises for quality enhancement.

In Section 4.2.5, we investigated the impact of the SNR of the recordings on the overall accuracy of the classification models. The results show that all classification models benefit from increased SNR levels, albeit with earlier plateauing for Mel-SVM and Log-CNN models. These results indicate that applying noise reduction techniques with the aim of improving the SNR of the recordings has a beneficial effect on the accuracy of the classification models. However, the level of performance gain varies from model to model.

Based on the above discussion, the implications of this study can be summarized as follows: first, in some cases, it is possible to reduce the negative impact of noise and degradation on the classification models at the heart sound acquisition stage. For example, in Section 4.2.1, we observed that classification models perform worse on short-duration heart sounds compared to long-duration ones. Also, in Section 4.2.3, we saw that ambient noises have a more detrimental impact on classification models' accuracy than internal or movement noises. By capturing long-duration signals or reducing ambient noises, clinicians will be able to decrease the destructive impact of such degradations on the performance of the classification models. Therefore, clinicians can use this study's results to adapt the heart sound capture process to minimize the negative impact of such noises and degradations. Second, in order to decrease the misclassification rate of classification models, it is necessary to assess the captured heart sound signals in terms of noise and degradation characteristics. For example, in Section 4.2.5, we observed that classification models show a higher misclassification rate when evaluated using the signals with lower SNR levels. Also, in Section 4.2.4, we saw that long-duration noises have a more destructive impact on the accuracy of the classification models compared to short-duration noises. Assessing the characteristics of heart sound signals, such as SNR at the pre-processing stage, allows us to discard low-quality heart sounds or adjust the quality enhancement based on the noise characteristics of the signal. Quality enhancement of the heart sound signals has been widely employed in the field as a pre-processing step to develop heart sound classification models [4,5,16,41]. However, quality enhancement algorithms have been universally applied to heart sound recordings, irrespective of the characteristics of noise and degradation in the signal. At the same time, it has been shown that universal quality enhancement can reduce the performance of classification models [42]. The results of this study show that the characteristics of noise and degradation in a heart sound recording determine how and to what extent the classification models are influenced. In this regard, assessing the characteristics of the noise and degradations in heart sound signals will allow us to develop *targeted* quality enhancement techniques which adapt the type and aggressiveness of quality enhancement depending on the noise content of the signals and the employed classification model.

5.3. Limitations

This study has some potential limitations. Firstly, as discussed in Section 3.1.1, we used 16 baseline heart sounds (8 normal, 8 abnormal) as the seeding instances to generate a dataset of 3360 samples. This dataset represents a controlled set of noise types and degradations to answer our specific research question. We should note that this study does not attempt to discover the impact of noise and degradation on the performance of the data-driven models under a wide variety of conditions. In the future, this research work can be repeated using a dataset generated with a larger variety of baseline heart sounds to

better understand how the size and scale of the dataset can affect the results. Also, this study explored the impact of noises and degradations on a small set of classification models. While a large variety of feature representations and classifiers have been employed in the field to develop heart sound classification models, in this study, we focused on the classification models used most frequently. However, there is still a slight possibility that other classification models react differently to noises and degradations in heart sound signals. Finally, as discussed in Section 3.2, we employed a segmentation-free heart sound classification pipeline, which means that we used fixed-length signals to train and evaluate the classification models. In other words, we did not apply segmentation algorithms to segment heart sound recordings into heartbeat cycles. Therefore, the results of this study may not be generalizable to the cases where segmentation algorithms are used as one of the stages in the modeling pipeline.

6. Conclusion

Noise and degradation in heart sound recordings can reduce the accuracy of the data-driven classification models. In this study, we investigated how and to what extent different heart sound signal characteristics such as signal duration, noise type, noise duration and SNR influence the performance of classification models. The general findings can be summarized as follows:

- The data-driven models perform worse on short-duration signals than long-duration ones. Therefore, we suggest using signals with at least 5-s durations when using a trained model for inference.
- Based on the results, color (white and pink) and continuous ambient noises (e.g., music and TV) are the most problematic noise sources for models. On the other hand, internal (e.g., cough and breathing), movement (body and sensor movement) and transient ambient noises (e.g., phone ring, dog bark) are less problematic for data-driven models.
- The classification models perform better on the signals contaminated with transient short-duration noises (e.g., cough, sensor movement and dog bark) than the ones with continuous long-duration noises (e.g., white or music noise). This is due to the fact that transient noises contaminate only a small portion of the signal while the remaining parts can have a high SNR as opposed to continuous noises that contaminate the entire signal uniformly. In other words, the results show that classification models can achieve good accuracy on signals with high *segmental SNR*, even if such signals have a low *average SNR*.

Although these findings may be intuitive from a human auditory perception perspective, this study quantifies them to a certain extent and guides us in terms of filtering heart sound data before presenting them to data-driven classification models. Clinicians can also use the findings of this study to identify noise and degradations that are more problematic to classification models and consequently adapt the heart sound capture process to reduce the negative impact of such degradations.

Comparing the findings of this study with the results of a survey we previously carried out with a group of clinicians regarding the impact of noise on the diagnosability of heart sounds indicates that clinicians and data-driven models suffer from noise and degradations in a similar manner. The survey results showed that from the point of view of respondents, short-duration heart sound recordings cannot be used towards diagnostics. Also, clinicians believed that ambient noises are more problematic to accurate diagnostics than internal and movement noises. Lastly, they stated that continuous long-duration noises are more disruptive to accurate diagnostics than transient short-duration noises. We can see a good agreement between the impact of *signal duration*, *noise types* and *noise duration* on the performance of classification models and diagnosability of heart sounds from the point

of view of clinicians. Therefore, the findings from our previous study are now backed up by the results of this study.

Universal heart sound quality enhancement, which has been frequently employed in the field as a pre-processing step, applies enhancement algorithms irrespective of the noise characteristics of the signals. However, the results of this study reinforce the importance of signal quality assessment in the heart sound classification pipelines. Quality assessment enables us to analyze the captured signals in terms of noise and degradations and limit the application of quality enhancement algorithms to specific signals which do not meet a required quality threshold. The findings of this study can be leveraged to develop targeted heart sound quality enhancement approaches which adapt the type and aggressiveness of quality enhancement based on the characteristics of noise and degradations in heart sound signals.

In future, we will extend this work by generating a larger synthetic dataset using a more diverse set of base heart sounds. Also, we will include a larger set of classification models to see if the results of this study hold for other heart sound classification models as well. As another future work, we will compare the universal and targeted heart sound quality enhancement approaches to determine which approach can better reduce the misclassification rate of heart sound classification models.

CRedit authorship contribution statement

Davoud Shariat Panah: Conceptualization, Methodology, Software, Data curation, Visualization, Writing – original draft. **Andrew Hines:** Supervision, Writing – review & editing. **Susan McKeever:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data available at <https://arrow.tudublin.ie/datas/20/>.

Acknowledgments

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. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- [1] Cardiovascular diseases, 2022, [Online; accessed 2022-10-02]. URL <https://www.who.int/health-topics/cardiovascular-diseases>.
- [2] J.A. Shaver, Cardiac auscultation: A cost-effective diagnostic skill, *Curr. Probl. Cardiol.* 20 (7) (1995) 447–530, [http://dx.doi.org/10.1016/S0146-2806\(07\)80002-8](http://dx.doi.org/10.1016/S0146-2806(07)80002-8), publisher: Mosby.
- [3] A.J. Taylor (Ed.), *Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation*, Springer-Verlag, London, 2015, <http://dx.doi.org/10.1007/978-1-4471-6738-9>, URL <https://www.springer.com/gp/book/9781447167372>.
- [4] C.N. Gupta, R. Palaniappan, S. Swaminathan, S.M. Krishnan, Neural network classification of homomorphic segmented heart sounds, *Appl. Soft Comput.* 7 (1) (2007) 286–297, <http://dx.doi.org/10.1016/j.asoc.2005.06.006>.
- [5] Z. Abduh, E.A. Nehary, M. Abdel Wahed, Y.M. Kadah, Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers, *Biomed. Signal Process. Control* 57 (2020) 101788, <http://dx.doi.org/10.1016/j.bspc.2019.101788>.
- [6] M. Alkhodari, L. Fraiwan, Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings, *Comput. Methods Programs Biomed.* 200 (2021) 105940, <http://dx.doi.org/10.1016/j.cmpb.2021.105940>.

- [7] P. Dhar, S. Dutta, V. Mukherjee, Cross-wavelet assisted convolution neural network (AlexNet) approach for phonocardiogram signals classification, *Biomed. Signal Process. Control* 63 (2021) 102142, <http://dx.doi.org/10.1016/j.bspc.2020.102142>.
- [8] M.B. Er, Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features, *Appl. Acoust.* 180 (2021) 108152, <http://dx.doi.org/10.1016/j.apacoust.2021.108152>.
- [9] Littmann electronic stethoscope model 3200, 2022, [Online; accessed 2022-10-01]. URL https://www.littmann.com/3M/en_US/littmann-stethoscopes/advantages/core-digital-stethoscope/.
- [10] CORE digital stethoscope - electronic stethoscopes | eko, 2022, [Online; accessed 2022-10-01]. URL <https://shop.ekohealth.com/products/core-digital-stethoscope>.
- [11] Jabes electronic stethoscope, 2022, [Online; accessed 2022-10-01]. URL <https://www.allheart.com/jabes-electronic-stethoscope/p/jsjabes3/>.
- [12] A.S. Paul, E.A. Wan, A.T. Nelson, Noise reduction for heart sounds using a modified minimum-mean squared error estimator with ECG gating, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2006, pp. 3385–3390.
- [13] D. Kumar, P. Carvalho, M. Antunes, R.P. Paiva, J. Henriques, Noise detection during heart sound recording using periodicity signatures, *Physiol. Meas.* 32 (5) (2011) 599–618, <http://dx.doi.org/10.1088/0967-3334/32/5/008>.
- [14] D. Gradolewski, G. Redlarski, Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment, *Comput. Biol. Med.* 52 (2014) 119–129, <http://dx.doi.org/10.1016/j.combiomed.2014.06.011>.
- [15] P.K. Jain, A.K. Tiwari, An adaptive thresholding method for the wavelet based denoising of phonocardiogram signal, *Biomed. Signal Process. Control* 38 (2017) 388–399, <http://dx.doi.org/10.1016/j.bspc.2017.07.002>.
- [16] P. Chen, Q. Zhang, Classification of heart sounds using discrete time-frequency energy feature based on S transform and the wavelet threshold denoising, *Biomed. Signal Process. Control* 57 (2020) 101684, <http://dx.doi.org/10.1016/j.bspc.2019.101684>.
- [17] S.R. Messer, J. Agzarian, D. Abbott, Optimal wavelet denoising for phonocardiograms, *Microelectron. J.* 32 (12) (2001) 931–941, [http://dx.doi.org/10.1016/S0026-2692\(01\)00095-7](http://dx.doi.org/10.1016/S0026-2692(01)00095-7).
- [18] D.B. Springer, T. Brennan, N. Ntusi, H.Y. Abdelrahman, L.J. Zühlke, B.M. Mayosi, L. Tarassenko, G.D. Clifford, Automated signal quality assessment of mobile phone-recorded heart sound signals, *J. Med. Eng. Technol.* 40 (7–8) (2016) 342–355, <http://dx.doi.org/10.1080/03091902.2016.1213902>.
- [19] H. Naseri, M.R. Homaeinezhad, Computerized quality assessment of phonocardiogram signal measurement-acquisition parameters, *J. Med. Eng. Technol.* 36 (6) (2012) 308–318, <http://dx.doi.org/10.3109/03091902.2012.684832>.
- [20] Q.-u.-A. Mubarak, M.U. Akram, A. Shaikat, F. Hussain, S.G. Khawaja, W.H. Butt, Analysis of PCG signals using quality assessment and homomorphic filters for localization and classification of heart sounds, *Comput. Methods Programs Biomed.* 164 (2018) 143–157, <http://dx.doi.org/10.1016/j.cmpb.2018.07.006>.
- [21] D. Shariat Panah, A. Hines, J.A. McKeever, S. McKeever, An audio processing pipeline for acquiring diagnostic quality heart sounds via mobile phone, *Comput. Biol. Med.* 145 (2022) 105415, <http://dx.doi.org/10.1016/j.combiomed.2022.105415>.
- [22] D.M. Shindler, Practical cardiac auscultation, *Crit. Care Nurs. Q.* 30 (2) (2007) 166–180, <http://dx.doi.org/10.1097/01.CNQ.0000264260.20994.36>.
- [23] J.S. Coviello, *Auscultation Skills: Breath & Heart Sounds*, Lippincott Williams & Wilkins, 2013.
- [24] N. Ranganathan, V. Sivacyan, F.B. Saksena, *The Art and Science of Cardiac Physical Examination*, JP Medical Ltd, 2015.
- [25] J. Li, R. Gadde, B. Ginsburg, V. Lavrukhin, Training neural speech recognition systems with synthetic speech augmentation, 2018, arXiv preprint [arXiv:1811.00707](https://arxiv.org/abs/1811.00707).
- [26] M. Severini, D. Ferretti, E. Principi, S. Squartini, Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation, *IEEE Access* 7 (2019) 51982–51993.
- [27] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, T. Vetter, Analyzing and reducing the damage of dataset bias to face recognition with synthetic data, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
- [28] V. Bolón-Canedo, N. Sánchez-Marño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, *Knowl. Inf. Syst.* 34 (3) (2013) 483–519, <http://dx.doi.org/10.1007/s10115-012-0487-8>.
- [29] D. Gradolewski, G. Magenes, S. Johansson, W. Kulesza, A wavelet transform-based neural network denoising algorithm for mobile phonocardiography, *Sensors* 19 (4) (2019) 957, <http://dx.doi.org/10.3390/s19040957>.
- [30] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, F. Castells, J.M. Roig, I. Silva, A.E.W. Johnson, Z. Syed, S.E. Schmidt, C.D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M.R. Samieinasab, R. Sameni, R.G. Mark, G.D. Clifford, An open access database for the evaluation of heart sound algorithms, *Physiol. Meas.* 37 (12) (2016) 2181–2213, <http://dx.doi.org/10.1088/0967-3334/37/12/2181>.
- [31] X. Bao, Y. Xu, H.-K. Lam, M. Trabelsi, I. Chihli, L. Sidhom, E.N. Kamavuako, Time-frequency distributions of heart sound signals: A comparative study using convolutional neural networks, 2022, [arXiv:2208.03128](https://arxiv.org/abs/2208.03128) [cs, eess]. URL <http://arxiv.org/abs/2208.03128>.
- [32] J. Allen, L. Rabiner, A unified approach to short-time Fourier analysis and synthesis, *Proc. IEEE* 65 (11) (1977) 1558–1564, <http://dx.doi.org/10.1109/PROC.1977.10770>, event-title: Proceedings of the IEEE.
- [33] D.D. O'Shaughnessy, *Speech Communications - Human and Machine*, second ed., 2000.
- [34] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, *Librosa: Audio and music signal analysis in python*, in: *Proceedings of the 14th Python in Science Conference*, Vol. 8, Citeseer, 2015, pp. 18–25.
- [35] A. Yadav, A. Singh, M.K. Dutta, C.M. Travieso, Machine learning-based classification of cardiac diseases from PCG recorded heart sounds, *Neural Comput. Appl.* (2019) <http://dx.doi.org/10.1007/s00521-019-04547-5>, [Online; accessed 2020-04-12]. URL <http://link.springer.com/10.1007/s00521-019-04547-5>.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in Python, in: *Machine Learning in Python*, 2011, p. 6.
- [37] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, <http://dx.doi.org/10.48550/arXiv.1412.6980>, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) [cs]. URL <http://arxiv.org/abs/1412.6980>.
- [38] Y. Chen, S. Wei, Y. Zhang, Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network, *Med. Biol. Eng. Comput.* 58 (9) (2020) 2039–2047, <http://dx.doi.org/10.1007/s11517-020-02218-5>.
- [39] B. Xiao, Y. Xu, X. Bi, J. Zhang, X. Ma, Heart sounds classification using a novel 1-D convolutional neural network with extremely low parameter consumption, *Neurocomputing* 392 (2020) 153–159, <http://dx.doi.org/10.1016/j.neucom.2018.09.101>.
- [40] T. Li, Y. Yin, K. Ma, S. Zhang, M. Liu, Lightweight end-to-end neural network model for automatic heart sound classification, *Information* 12 (2) (2021) 54, <http://dx.doi.org/10.3390/info12020054>.
- [41] T.H. Chowdhury, K.N. Poudel, Y. Hu, Time-frequency analysis, denoising, compression, segmentation, and classification of PCG signals, *IEEE Access* 8 (2020) 160882–160890, <http://dx.doi.org/10.1109/ACCESS.2020.3020806>.
- [42] M.H. Asmare, F. Woldehanna, L. Janssens, B. Vanrumste, Can heart sound denoising be beneficial in phonocardiogram classification tasks?, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2021, pp. 354–358.