Alexandria Engineering Journal (2022) 61, 10995–11002

HOSTED BY

ELSEVIER

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

Heartbeat murmurs detection in phonocardiogram recordings via transfer learning

Omair Rashed Abdulwareth Almanifi [a], Ahmad Fakhri Ab Nasir [a], Mohd Azraai Mohd Razman [a], Rabiu Muazu Musa [b], Anwar P.P. Abdul Majeed [a,c,d,*]

[a] *Innovative Manufacturing, Mechatronics and Sports Laboratory, Faculty of Manufacturing and Mechatronics Engineering Technology, Universiti Malaysia Pahang (UMP), 26600 Pekan, Pahang Darul Makmur, Malaysia*
[b] *Centre for Fundamental and Liberal Education, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu Darul Iman, Malaysia*
[c] *School of Robotics, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, 215123, P. R. China*
[d] *EUREKA Robotics Centre, Cardiff School of Technologies, Cardiff Metropolitan University, CF5 2YB, Cardiff, UK*

Received 30 September 2020; revised 7 April 2021; accepted 16 April 2022

KEYWORDS

Transfer learning;
Convolution neural networks;
Phonocardiogram;
Spectrograms;
Mel frequency cepstral coefficients

Abstract Heart murmurs are abnormal heartbeat patterns that could be indicative of a serious heart condition, which can only be detected by trained specialists with the use of a stethoscope. However, it is occasionally the case that those specialists are not available, resulting in the need for a machine-automated system for murmur detection. Many methods might be used to produce such a system, one of which is the utilization of transfer learning. A recent machine learning method that saw popularity due to the little time it needs for training and the boosted accuracy it produces. This paper aims at testing the performance of transfer learning when detecting murmurs of the heart, by evaluating three transfer learning models, namely, VGG16, VGG19, and ResNet50, trained on a database of phonocardiogram (PCG) heartbeat recordings, i.e., PASCAL CHSC database. The data is cleansed, processed, and converted into images using two signal representation methods; Spectrograms and Mel Frequency Cepstral Coefficients (MFCCs). The paper compares the results of each model, using metrics of accuracy and loss, where the use of Spectrograms proved to yield the best results with 83.95%, 83.95%, and 87.65%, classification accuracy for VGG16, VGG19, and ResNet50, respectively. Based on the findings of the paper, it is evident that the Spectrogram-ResNet50 transfer learning pipeline could further facilitate the detection of heart murmurs with less time spent on training.

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally. In 2016 alone 17.9 million people died of CVD, making 31% of the overall number of fatalities [1]. More than 70%

* Corresponding author at: School of Robotics, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University. Suzhou. 215123. P. R. China.
E-mail address: anwarmajeed1983@gmail.com (A. P.P. Abdul Majeed).

Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2022.04.031
1110-0168

of all CVD deaths are in countries with low and middle income. Most diseases of this type can be prevented, and early detection of risk is vital to the process of prevention. Heart murmur in many cases is an early indicator of CVDs. The sound pattern that murmur makes is normally perceived by doctors and physicians using a stethoscope in a process called Cardiovascular auscultation [2].

Normal hearts make a very similar pattern resultant from the closing and opening of heart valves, which can be observed in healthy humans with no threatening heart conditions. Any divergence from this pattern is called an abnormality, and murmur is an example of that. Heart murmurs might not always be harmful; nonetheless, they could be an indicator of several serious heart problems [3]. Trained doctors can easily pick up on the sound of murmurs; however, it is not always the case that an expert doctor is available, especially in rural areas suffering from a shortage of specialist medical doctors.

In a newly introduced solution that utilizes machine learning, heartbeat signals are recorded in a digital form called phonocardiogram (PCG) [4]. PCG sound signals are then processed for feature extraction, which aids in the process of classification. There are many techniques applied when it comes to feature extraction and many architectures designed for classification. Each and every one of these yields different results when it comes to accuracy and loss.

One of the many applications of machine learning for the detection of murmur was by Yaseen et al. [5]. A sample of 1000 signals, 800 normal and 200 abnormal signals, was used, with three different techniques of feature extraction, namely Mel Frequency Cepstral Coefficients (MFCCs), Discrete Wavelets Transform (DWT), and the combination of both. Support vector machine (SVM), k-nearest neighbors (KNN), and Deep Neural Network (DNN) are the machine learning models used and compared in the paper. The highest accuracy achieved was 97.9% when using SVM with both MFCCs and DWT.

Ahmad et al. [6], used 283 sound samples, recorded at the cardiology department of Ayub Teaching Hospital, Abbottabad, Pakistan. The chosen ML models were SVM, and KNN, with MFCCs as the main feature extraction method. With the use of five-fold cross-validation and 20% data hold-out as the cross-validation techniques. An accuracy of 92.6% was achieved by the SVM model.

Another application was demonstrated in a study by Lubis et al.[7]; using back-propagation neural networks for the detection of murmur. The data used was taken from The Pascal Classifying Heart Sound database, namely dataset B, where DCT, MFCC – and BFCC were used for feature extraction and k-fold cross-validation as the main cross-validation technique. The study concluded that the use of MFCC yielded an average cross-validated classification accuracy of 63.54%, compared to modified MFCC with an average accuracy of 61.45%. These relatively low results are the results of the limited, noisy dataset.

A self-produced wireless electric stethoscope was produced in a study conducted by Choi et al.[8], where a multi-SVM was deployed for the classification process. The dataset used in this study consisted of 489 cardiac sound signals, 196 of which were normal, and 293 abnormal cases, all acquired from 34 patients. As for the feature extraction methods used in the study, a two-dimensional representation was produced by illustrating the maximum peaks of the normalized AR-PSD curve on a selected threshold value defining two morphological feature values Fmax and Fwidth. The results in this study were dependent on the threshold, as a THV of 10–90% gives an average accuracy of 88.4%, and an average accuracy of 90% with THVs of 10–50%.

Recently a new machine learning (ML) technique found its way into the spotlight as a way of significantly increasing the accuracy of ML models. This technique is called, transfer learning; a way of transferring the weights of a pre-trained model to a new model, thus increasing its accuracy. The models transferred typically need to be trained on a very wide data-set and must be able to classify many classes as well to be useful. Many of these models were developed, such as VGG, ResNet, and Inception, and they are used to boost the performance of ML models in all types of fields. Though there are no studies on the performance of transfer learning for PCG classification, it has been used widely in the field of biomedicine.

Mahbod et al. [9] conducted a recent study that is a prime example of that; using transfer learning for skin lesion classification. The transferred models were EfficientNetB0, EfficientNetB1, and SeReNeXt-50, which are pre-trained CNNs, they were used individually and as multi-scale networks. They were trained on the skin lesion classification challenge datasets, provided by International Skin Imaging Collaboration (ISIC), producing promising results with balanced multi-class accuracy of 86.2%.

Another instance for the use of transfer learning is in a study by Steenkiste et al.[10], where transfer learning was used to classify ECG from humans to horses. The study was conducted on a self-produced eECG data set, and the MIT-BIH arrhythmia dataset. Using wavelet transforms the authors compared the performance of a parallel CNN architecture with and without transfer learning where it recorded an accuracy of 92.6% without transfer learning and an accuracy of 97.1% with the use of transfer learning.

It can be noted that there are many factors that control the performance of ML models on a general scale [11,12], and there exist many performance-enhancing techniques to be tested and studied [13]. Due to its proven success in aiding with the training of small datasets and its capability to reduce the training time, the use of transfer learning for the detection of heart murmur is the main focus of this paper. Using the publicly available PASCAL CHSC 2011 database [14], pre-trained Convolutional neural networks (CNN) models, namely VGG16, VGG19, and ResNet50, were tested in the study, with the utilization of Spectrogram and MCFFs signal representation methods.

2. Materials and methods

As mentioned above, PCG recordings can be very useful in detecting murmur in the human heart. They are represented digitally by audio signals with different frequency range depending on the method of recording and machine used, this signal can be processed to aid with the task of detection. A dataset should be acquired, the signals out to be cleaned from noise and sampled at a specific frequency rate. The clean signal is then ready for feature extraction, before being classified into one of the specified classes using the chosen machine learning, or to a certain extent, deep learning models after being developed, trained, and tested.

2.1. Data acquisition

When it comes to PCG datasets, there are a few available online for public use. One of these is the PASCAL CHSC 2011 database [14], which is used in this paper mainly owing to the fact that it is inclusive of non-heartbeat audio tracks, and tracks sampled without a professional setting, using an IOS application. This is to test bring about the full potential of transfer learning as it can give a relatively great performance even with small and noisy datasets. The databased was acquired from two sources: healthy non-patients from the general public through the iStethoscope Pro iPhone app (Section A), and from hospitals using DigiScope, a digital stethoscope (Section B). The dataset consists of 5 categories: Artifacts, which are non-heartbeat sounds, normal heartbeats, murmur, extrasystole, and extrahls (normal heartbeats with additional sounds) all in WAV format. The total number of tracks in the dataset is 832, however, after cleaning unusable tracks (audio that is less than 5 s, and unplayable tracks), the number that was finally used in the present study is 404.

2.2. Feature extraction

Signal representation techniques are many in the field of signal processing, and they are a very useful tool for the analysis of signals. For the purpose of this study, they are used to extract classifiable features. Primary examples of such techniques are Spectrograms and Mel-frequency cepstral coefficients (MFCCs), which are used in this study, primarily because of their popularity and flexibility. Spectrograms are used in a wide range of fields pertaining to signal analysis since it is a basic and flexible method of signal visualization [15]. MFCCs, on the other hand, are used widely in speech analysis due to their quasi-logarithmic spacing which bears some resemblance to the human auditory system [16]. MFCCs represent the power spectrum of sound, transforming it using linear cosine transform. Fig. 1 shows the derivation method for MFCCs. A comparison between the audio signal, a spectrogram, and MFCCs is illustrated in Fig. 2.

In the present investigation, the Librosa [17] python package was used to convert the audio signals to Spectrogram and MFCC representations. The package was as well used for noise cleansing and sampling as it is equipped with many available tools for this purpose. The minimum frequency used was 0, and the maximum was half of the sampling rate, which is 22050 Hz. The number of mel bins is 96 and the length of the windowed signal padding with zeros is 2048 samples in the spectrograms. As for MFCCs, 40 amplitudes of the Discrete Cosine Transform (DCT) spectrums were used.

2.3. Classification

Recently, transfer learning was introduced as a way of storing and transferring knowledge between machine learning models for the goal of increasing the overall accuracy of the latter [18]. When ML models are trained, the weights and biases change to reach the highest accuracy the model can achieve, the weights are then saved to be used to make predictions [19]. In transfer learning, the weights of pre-trained Convolutional Neural Networks (CNN) models are transferred to a new model. The transferred models can be used solely, however, it will yield bad results as the model will not acquire any new knowledge. Therefore, some layers must be added for training. Fig. 3 demonstrates the process of transfer learning. Note that the model is separated from the classification layer and the dataset input, before being moved frozen to the new model.

In the present study, three models are compared: VGG 16 [20], a model developed at the University of Oxford, which uses 16 layers of convolution and pooling, with multiple kernels of size 3*3. VGG19, which is an improvement on the previous model with 19 layers instead of 16 [21]. Lastly, ResNet50, short for Residual Networks, is a 50 layered model that uses the concept of skipping layers, which allows the model to skip layers at times, preventing overfitting, solving the problem of vanishing gradient, and helping higher layers perform as good as lower ones [22]. All these models were trained on ImageNet [23], a dataset made for object detection research containing more than 14 million images.

Python [24], a high-level general-purpose computer language, was used to process the data, and develop the ML model. The language was chosen for its flexibility, its libraries' diversity, its easiness to use, and straightforwardness. Several libraries and packages were used, such as Librosa as mentioned above, Numpy [25], a library for scientific computing, Matplotlib [26], a visualization library used for graphing and visualizing data, Pandas [27], a package that deals with CSV files, Keras [28], a package that uses Tensorflow backend [29], it facilitates the development of machine learning algorithms and models, and lastly, Scikit-learn [30], which was used for splitting data, and measuring accuracy. The development of the model initiated with transferring the previously mentioned models, which were then connected to a 20% dropout layer and a dense layer with 64 parameters, activated with a SeLU activation function. Finally, the model is connected to the classification dense layer activated with Softmax. Note that the same structure was used for all models to produce a fair comparison. After the development of the models, they were trained on the datasets with an 80:20 training–testing splitting ratio.

2.4. Performance evaluation

There are many metrics for the assessment of the performance of an ML model. In this study "accuracy" and "loss" are the metrics used to monitor the performance of the models. Accuracy in this context is a measure of the percentage of correct predictions from the total number of data made by the model, it is obtained by:

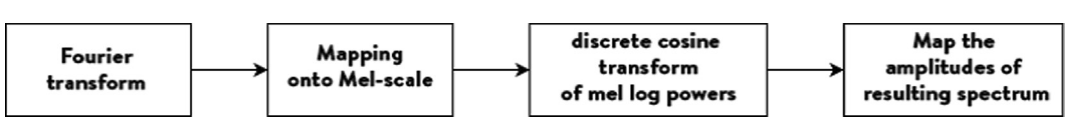

Fig. 1 Flow of the derivation method for MFCC.

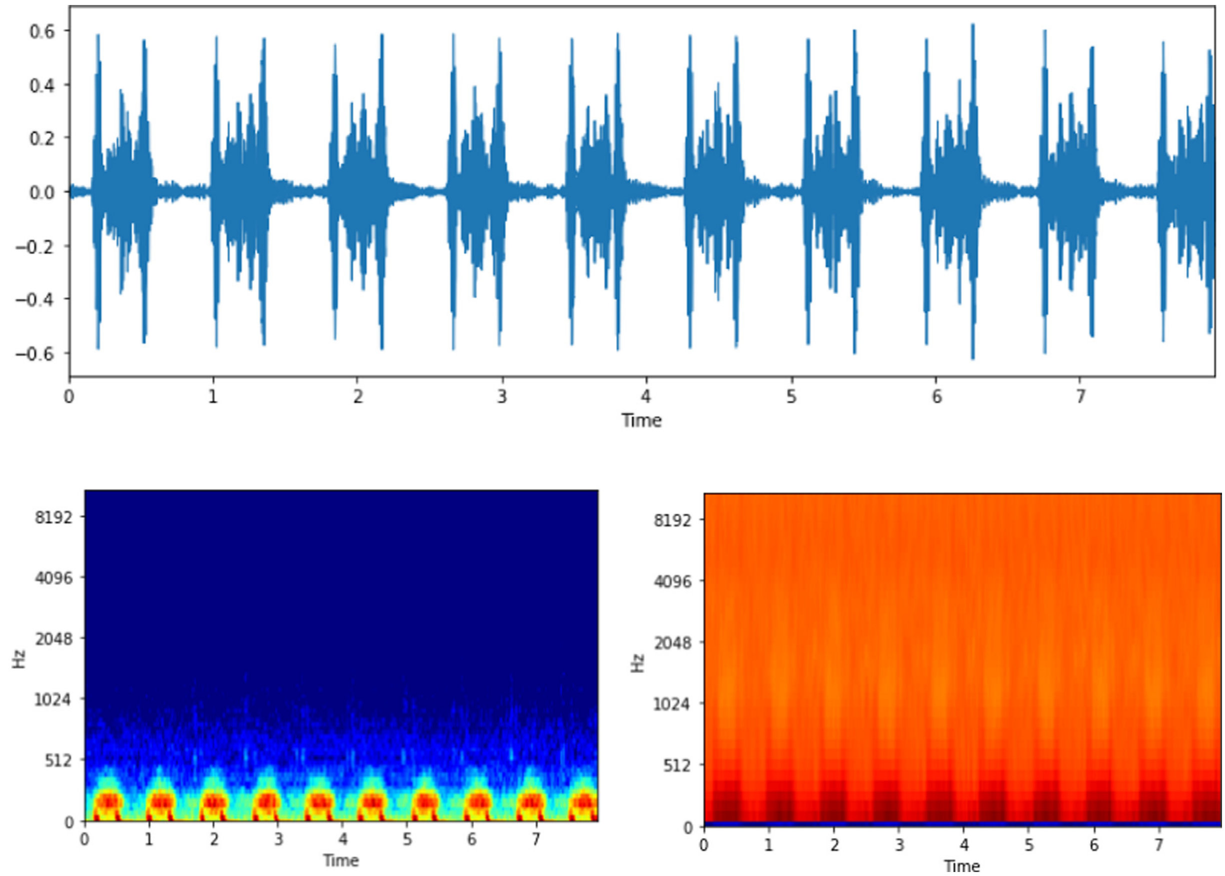

Fig. 2 Spectrogram (bottom left) and MFCC (bottom right) representation of the audio signal.

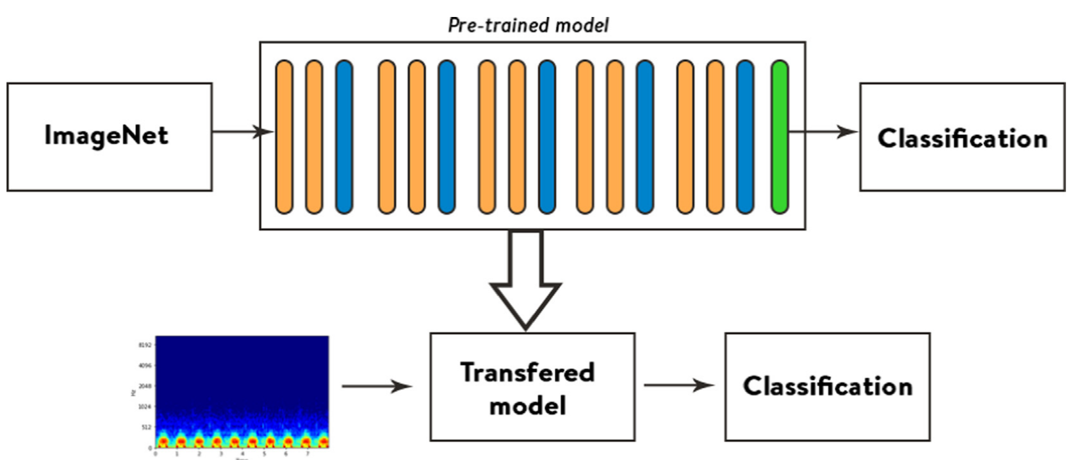

Fig. 3 Transfer Learning methodology.

$$Accuracy = \frac{N_c}{N_t}$$

where N_c is the number of correct predictions made by the model, and N_t being the total number of true data.

As for loss, it is defined as the difference between the model predictions and the original data. There are many loss functions, the one discussed here is cross-entropy, since we have 5 different classes, and it can be defined as the sum of errors made by the model during the process of training of testing [31]. It is given by:

$$Loss_{\text{cross-entropy}} = \sum_{k=1} \sum_{l=1} y_{p(k,l)} \log(p_{(k,l)})$$

where y_p is the classification choice, where it is 1 if the prediction (k) is accurate for each class (l) and 0 if otherwise, and p is the probability of the prediction being accurate, that is to say, it belongs to the class chosen.

3. Results and discussion

This paper uses the metrics of accuracy and loss mentioned above to compare three transfer learning models and signal visualization methods. Table 1, Fig. 4, and Fig. 5, demonstrate the classification accuracy and the loss values obtained by the model, where both training and testing sets are compared. Aside from the performance metrics, Table 1 includes the number of epochs each of the models needed before achieving their best performance and the amount of time it takes for each epoch, added to show the speed of training that could be achieved with transfer learning. The number of epochs was determined by utilizing early stopping with a patience value of 5 epochs. The three models are grouped by the used signal representation method to best illustrate the results found by the study.

Confusion matrix (also known as error matrix) is a visualization tool for the assessment of supervised learning models [32]. It maps out the predictions made by the model, where the rows represent the actual class of the data, and the columns the predicted class. It is normally produced so as to make the analysis of models more comprehensive, telling where what classes the model was best at predicting and what classes it was most liable to make errors predicting. Fig. 6 depicts six confusion matrices for all the models and feature extraction methods on the test dataset.

Studies have been done on the performance of CNN models for the classification of PCG recordings; however, the use of transfer learning remains unexplored. Thus, this study evaluates the performance of such techniques with respect to the pre-trained models and the method of feature extraction, which are VGG16, VGG19, and ResNet50 models, and spectrogram and MFCCs feature extraction methods.

The performance of the suggested models can be observed in Table 1, Fig. 4 and Fig. 5, where it can be noted that generally, the performance of the models when Spectrogram is used to visually represent the signals is greater, compared to its counterpart, MFCCs, which could be explained to be attributed to its sensitivity towards lower frequencies [33]. Observing the models individually, we can note that ResNet50 seems to be performing the best overall, with a classification accuracy of 87.65%, in comparison to VGG16 and VGG19 which scored similar results of 83.95% testing accuracy. As for the cross-entropy loss, all models seem to have a similar range with ResNet50 being slightly higher during testing.

In Table 1 the run time of the models was recorded in terms of epochs and the seconds per epochs. It can be observed that ResNet50 is the fastest at training, running only for 12 epochs, each taking only 6 s to complete when trained with Spectrogram, compared to VGG16 and VGG19 which took 35 and 55 epochs respectively, with a run time of 7 s and 8 s. It can be noted that when training with MFCC data, the number of epochs increases in the case of ResNet50 and VGG19 at 17 and 60 epochs, respectively, while VGG16 ran for only 33 epochs. This relatively small number of training epochs is due to the use of per-trained data, which can save a large portion of training time, a primary advantage of using transfer learning.

It is important to note that the difference between the training and testing accuracy and loss was due to the limitation of the database as it only has 404 audio tracks, which is visible in the confusion matrices in Fig. 6. Most accurate predictions were made by the class with the highest number of data (normal heartbeat), which indicates that the small number of training data was the cause behind the underperformance in testing as the model is unable to generalize. However, the performance is still relatively great in comparison with other studies that used the same dataset, such as the one conducted by Lubis et al [7] where the average classification accuracy was 63.54%. It is deduced that transfer learning can be used to boost the performance of noisy data, with less training time, however, further studies performed on a larger dataset could possibly yield better results.

4. Conclusion

To sum up, the present study tested the possibility of the use of transfer learning for the detection of heart murmur using PCG recordings. Using the PASCAL CHSC 2011 database, the signals were processed and sampled, and visually represented using spectrogram and MFCCs. Three pre-trained CNN models were used for classification namely VGG16, VGG19, and ResNet50, and the results indicate that the use of transfer learning for the task is not only possible but gives relatively great performance for the detection of murmurs, with the Spectrogram-ResNet50 pipeline achieving a classification accuracy of 87.65%. While the results were significantly improved compared to previous work done on the same dataset, there is still a large space for improvement; different machine learning techniques can be tested, such as ensemble, and a larger dataset could provide a great boost in performance in comparison to the limited one used in the study.

Table 1 The resulted training and test accuracy (%) and loss for each feature extraction method, and each transfer learning model.

Dataset	Metric	Feature extraction method					
		Spectrogram			MFCCs		
		VGG16	VGG19	RESNET50	VGG16	VGG19	RESNET50
Training	Accuracy	82.35%	86.07%	90.09%	81.11%	80.19%	90.04%
	Loss	0.8594	0.9126	0.9060	0.8181	0.7969	0.7431
Testing	Accuracy	83.95%	83.95%	87.65%	75.31%	77.78%	80.25%
	Loss	1.0204	0.9126	1.0825	0.9887	0.8472	1.0225
Run time	No. epochs	35	55	12	33	60	17
	Seconds per epoch	7 s	8 s	6 s	7 s	8 s	6 s

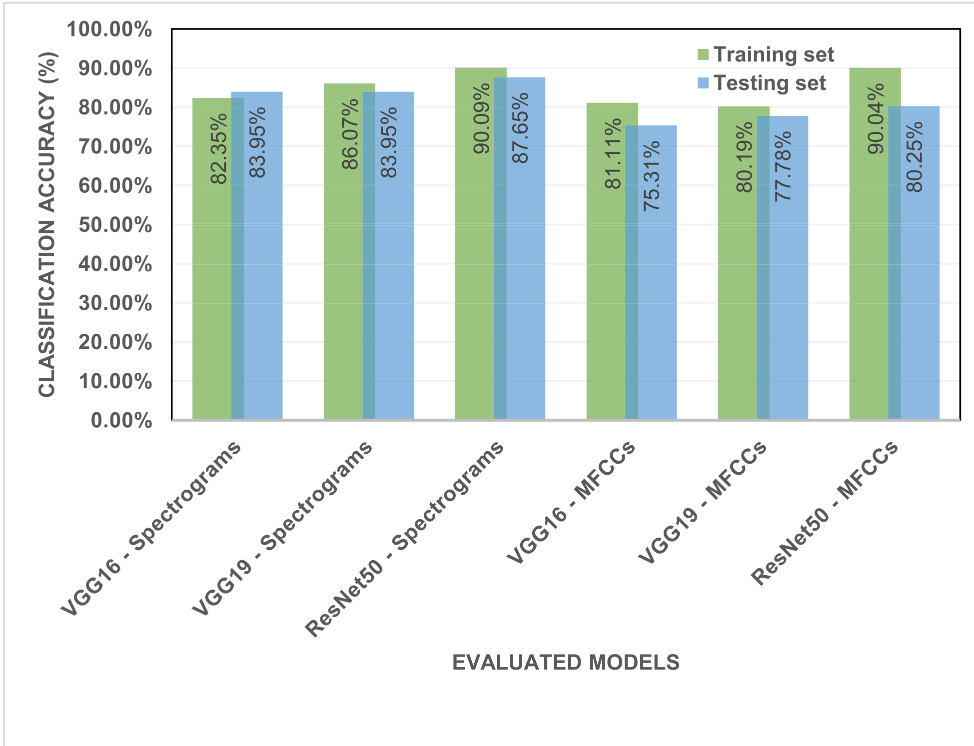

Fig. 4 The classification accuracy achieved by each model.

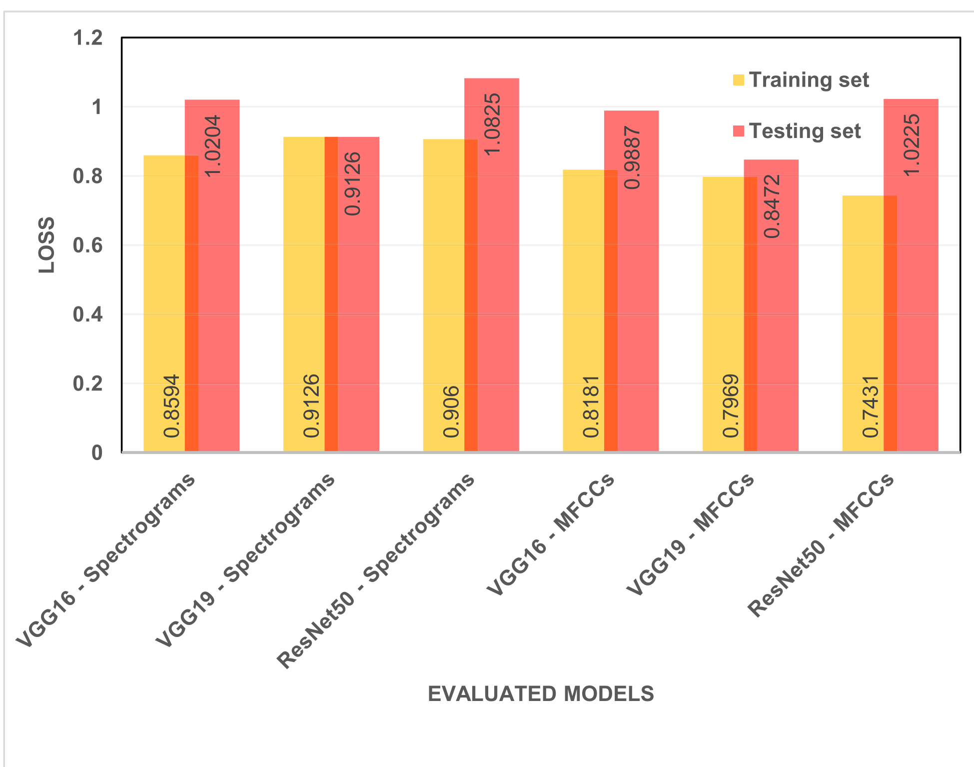

Fig. 5 The loss recorded by each model.

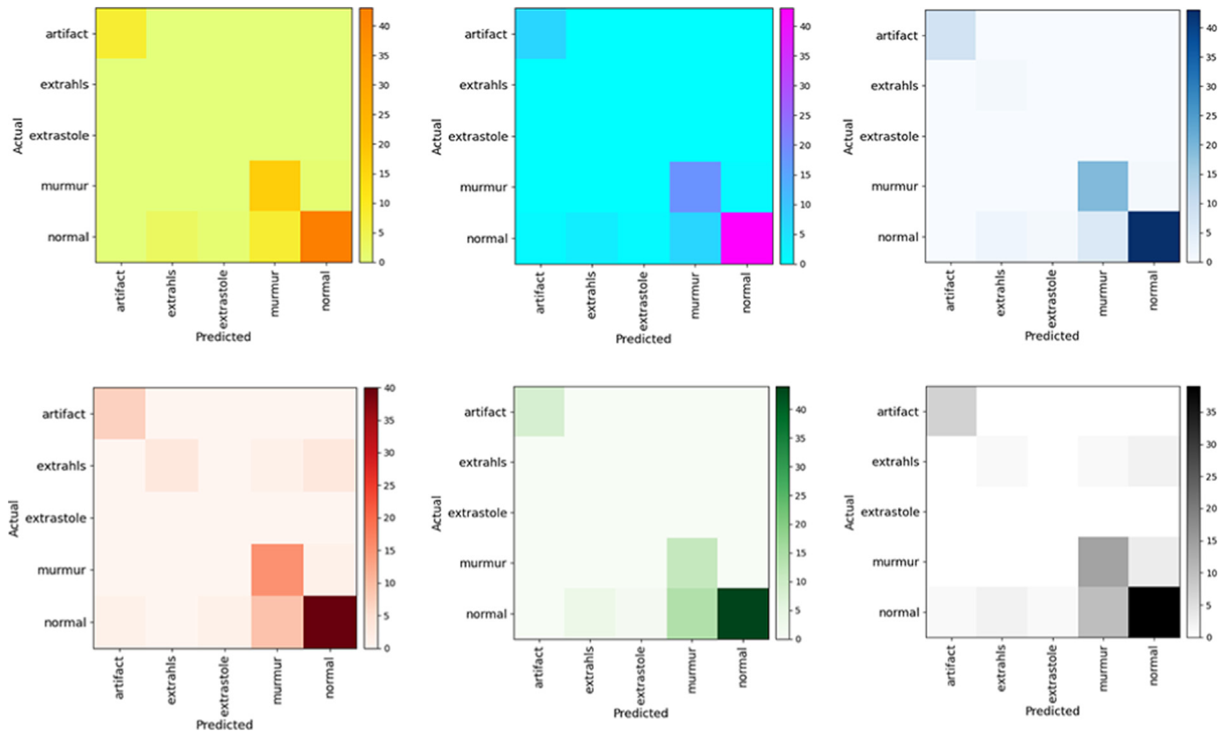

Fig. 6 Confusion matrices for each model.

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.

References

[1] WHO. Cardiovascular diseases (CVDs) fact sheets. Who. 2016.
[2] M.A. Chizner, Cardiac Auscultation: Rediscovering the Lost Art, Curr. Probl. Cardiol. (2008).
[3] Murmurs and Heart Disease. Encycl Hear Dis. 2011.
[4] G. Goodman, Cardiovascular Techniques and Technology, Clin. Eng. Handb. (2019).
[5] S.GY. Yaseen, S. Kwon, Classification of heart sound signal using multiple features, Appl. Sci. (2018).
[6] M.S. Ahmad, J. Mir, M.O. Ullah, M.L.U.R. Shahid, M.A. Syed, An efficient heart murmur recognition and cardiovascular disorders classification system, Australas. Phys. Eng. Sci. Med. (2019).
[7] C. Lubis, F. Gondawijaya, Heart Sound Diagnose System with BFCC, MFCC, and Backpropagation Neural Network, IOP Conf. Ser. Mater. Sci. Eng. (2019).
[8] S. Choi, Z. Jiang, Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique, Comput. Biol. Med. (2010).
[9] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, I. Ellinger, Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification, Comput. Methods Programs Biomed. (2020).
[10] G. Van Steenkiste, G. van Loon, G. Crevecoeur, Transfer Learning in ECG Classification from Human to Horse Using a Novel Parallel Neural Network Architecture, Sci. Rep. (2020).
[11] R. Iqbal, T. Maniak, F. Doctor, C. Karyotis, Fault Detection and Isolation in Industrial Processes Using Deep Learning Approaches, IEEE Trans. Ind. Informat. (2019).
[12] K. Iqbal, M. Odetayo, A. James, R. Iqbal, N. Kumar, S. Barma, An efficient image retrieval scheme for colour enhancement of embedded and distributed surveillance images, Neurocomputing (2016).
[13] R. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, Big Data analytics and Computational Intelligence for Cyber-Physical Systems: Recent trends and state of the art applications, Futur. Gener. Comput. Syst. (2020).
[14] E.F. Gomes, P.J. Bentley, M. Coimbra, E. Pereira, Y. Deng, Classifying heart sounds: Approaches to the PASCAL challenge, Heal 2013 – Proc. Int. Conf. Heal Informatics, 2013.
[15] W. Zhang, J. Han, S. Deng, Heart sound classification based on scaled spectrogram and partial least squares regression, in: Biomed Signal Process Control, Elsevier, 2017, pp. 20–28.
[16] V. Tiwari, MFCC and its applications in speaker recognition, Int. J. Emerg. Technol. (2010).
[17] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, et al, librosa: Audio and Music Signal Analysis in Python, Proc 14th Python Sci Conf., 2015.
[18] W. Transfer, L. Now, T.L. Scenarios, T.L. Methods, Transfer Learning - Machine Learning's Next Frontier, PPT. (2017).
[19] F.F. Qureshi, R. Iqbal, M. Qasim, F. Doctor, V. Chang, Integration of OMNI channels and machine learning with smart technologies, J. Ambient. Intell. Humaniz. Comput. (2017).
[20] Muneeb ul Hassan, VGG16 - Convolutional Network for Classification and Detection, Neurohive. (2018).
[21] M. Shaha, M. Pawar, Transfer Learning for Image Classification, Proc. 2nd Int. Conf. Electron Commun. Aerosp. Technol. ICECA 2018, 2018.
[22] N. Sharma, V. Jain, A. Mishra, An Analysis of Convolutional Neural Networks for Image Classification, Procedia Comput. Sci. (2018).

[23] A. Krizhevsky, I. Sutskever, E.H. Geoffrey, Imagenet, Adv. Neural. Inf. Process Syst. 25 (2012).
[24] N. Ketkar, Deep Learning with Python, Deep Learn. Python. (2017).
[25] NumPy. NumPy — NumPy. NumPy Website, 2017.
[26] C. Hill, H.C. Matplotlib, Learn Sci. Program. Python. (2016).
[27] W. McKinney, P.D. Team, Pandas - Powerful Python Data Analysis Toolkit, Pandas - Powerful Python Data Anal Toolkit. 2015.
[28] F. Chollet, Keras Documentation. Keras.Io. 2015.
[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, et al, TensorFlow: A system for large-scale machine learning, Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, 2016..
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. (2011).
[31] R.E. Neapolitan, R.E. Neapolitan, Neural Networks and Deep Learning, Artif. Intell. (2018).
[32] K.M. Ting, Confusion Matrix, Encycl Mach Learn Data Min, 2017.
[33] M. Mulimani, S.G. Koolagudi, Segmentation and characterization of acoustic event spectrograms using singular value decomposition, Expert Syst. Appl. 120 (2019) 413–425.