



Heart Sound Classification Using Multi Modal Data Representation and Deep Learning

Jang Hyung Lee¹, Sunyoung Kyung², Byungchun Oh³, Kwang Gi Kim^{1,*}, and Dongjin Shin⁴

¹Department of Biomedical Engineering, Gachon University School of Medicine, Incheon, South Korea

²Department of Internal Medicine, Gachon University Gil Medical Center, Incheon, South Korea

³Department of Cardiology, Gachon University Gil Medical Center, Incheon, South Korea

⁴Department of Neurology, Gachon University Gil Medical Center, Incheon, South Korea

Heart anomalies are an important class of medical conditions from personal, public health and social perspectives and hence accurate and timely diagnoses are important. Heartbeat features two well known amplitude peaks termed S1 and S2. Some sound classification models rely on segmented sound intervals referenced to the locations of detected S1 and S2 peaks, which are often missing due to physiological causes and/or artifacts from sound sampling process. The constituent and combined models we propose are free from segmentation, which consequently is more robust and meritorious from reliability aspects. Intuitive phonocardiogram representation with relatively simple deep learning architecture was found to be effective for classifying normal and abnormal heart sounds. A frequency spectrum based deep learning network also produced competitive classification results. When the classification models were merged in one via SVM, performance was seen to improve further. The SVM classification model, comprised of two time domain submodels and a frequency domain submodel, produced 0.9175 sensitivity, 0.8886 specificity and 0.9012 accuracy.

Keywords:

1. INTRODUCTION

Heart problems, being a leading cause of adult diseases, have deep implications over multitude of domains ranging from personal, social to public health. Accurate and accessible diagnosis is an essential first step towards the treatment of heart problems. While advanced diagnostic means have been introduced and are in use for detection of heart abnormalities, traditional stethoscopic method still provides for a handy and relatively simple way, particularly useful in circumstances including emergency [1]. Heart sound signal features structural peaks formed by the clapping motions of heart valves at the time of their closure and trailing smaller amplitude chirps from turbulent blood flows in heart chamber and arteries. A number of factors may degrade the quality of heart sound signals—including relatively low level of reliability of sound sampling and recording processes, noise stemming from various kinds of sources including friction between stethoscope and skin and dialog between medical staff and patients during diagnostic session [2]. Closing of mitral and tricuspid valves generates the S1 sound and the closing of arterial valves generates the S2 sound. Normal and healthy heart sounds overall tend to assume a degree of uniformity where the presence of the S1 and S2 peaks are main discernable features

with varying degree of perturbations from noise from various sources. Differences in sound across healthy individuals are due to variations in the separation between the S1 and S2 in time, possible dual peaks split off the peaks due to minor differences in the timing of closure of each valve [3, 4]. Noise and artifacts gathered during typical auscultation session tend to far outweigh the slight physiological variations across individuals to render normal heart sound signals quite varying overall. Heart problems can manifest in a variety of sound patterns—such as ejection click, early systolic murmur, late systolic murmur, opening snap and pan systolic murmur, etc., resulting from a range of pathological causes including sclerosis, stenosis and cardiomyopathy. In combination with the auscultation and recording artifacts, the abnormal sounds overall may assume great diversity [5]. Diagnostic decisions involving auditory data collected from stethoscope tend to be rather subjective whereby reproducibility, quantitateness and objectiveness are often hampered. Deep learning methods are expanding the horizons of computational means-based automatic cognitive and inferential mimetic processes, previously set forth by traditional machine learning methods [6]. We investigate ways for sound signal representation amenable to application of deep learning methods, develop automatic diagnosis methods using them and report an excellent performance result to bear clinical pertinence. Deep learning technique finds increasing uses in diverse application areas due to its capability of data-driven

* Author to whom correspondence should be addressed.

discovery for optimal representation under given data characteristics and problem solving objectives. Particularly in cognitive tasks involving images including classification, regression and clustering, substantial improvement over traditional methods has been brought by deep learning. Sound data is one dimensional in nature spanning over time. For the task of anomaly classification, different representations for modeling heart sounds are conceivable—such as hidden markov model and recurrent neural network [7, 8]. On the other hand, phonocardiogram plots out amplitudes of sound signal versus time in two dimensional plane as an image. The proverb “a picture is worth thousand words” implies visual representations are far more effective than raw phonetic or vocal representations in conveying information in some domains of cognitive tasks and problems. Characteristics spots of pathological signals may well be defined in relations to other characteristic signals including S1 and/or S2 or others. Discerning signals for diagnosis may appear in sporadic manners or span multiple periods of heart beats. Then listening to heart sound for diagnosis purpose implies localized access to signals which leaves the task of aggregating signals and synthesizing and transforming them towards decisions. On the other hand, phonocardiogram provides for a visual representation for sound signals possibly spanning multiple heart beat periods in a more comprehensive manner. Hence visual cardiogram has certain advantages in terms of economy of representation.

Models for deep learning typically entails a large number of parameters, values of which have to be appropriately set by using commensurate size of training data. It can encompass multiple heartbeat periods in a more comprehensive manner than one dimensional representation. By employing phonocardiogram representation and deep leaning together, rich resources in data, collective experiences and pre-trained weight parameter setting for models from research and development community can be utilized. Many prior works attempted to locate the heart sound peaks and to build models assuming such peaks available [9, 10]. Yet peaks are often unclear or absent from many data instances due to various artifacts, which diminishes the robustness of heart sound classification models relying on their presences. Dispensing with heart sound segmentation greatly reduces computational complexity, processing flow overall gets more streamlined and task can be performed in a manner closer to end to end fashion. While sound signal primarily resides in time domain, a dual form in frequency domain is well known to reveal richer aspects that are often less apparent in the time domain and has long been utilized for more thorough depiction of signal. Hence, we employed and tested both representations to build classification models in our study. Also composite ensemble models made up of models employing representations from the respective time and frequency domains were built so as to take advantage of any complementarity between them.

2. MATERIAL AND METHODS

Physionet challenge dataset is a comprehensive heart sound data set totaling 665 abnormal and 2575 normal samples [2]. Independently compiled sound data sets from multiple medical agencies were resampled to a uniform sampling frequency of 2kHz and pooled together to build the set. Other than resampling, no additional measures of processing such as noise removal, etc., were used to curate the comprising data instances. Six subsets are labeled a to f , and each subset exhibits distinct characteristics in

terms of sound quality, level of noise floor, etc., possibly due to differences in recording protocols used by the producing agencies. Since a primary consideration of the physionet challenge was to build a robust classification model capable of tolerating the variations in the sound quality and in recording protocols, we did not exercise any subset inclusion or exclusion criteria in forming train or test sets. For the purpose of building sound classification models using available data, we partitioned abnormal sets into subsets of sizes 265 and 400 respectively in a mutually exclusive manner to use as parts of train and test sets. Equal numbers of data were randomly selected from the normal set and pooled together with abnormal subsets to form train and test sets of respective sizes of 530 and 800. Both train and test sets were balanced in terms of comprising normal and abnormal data counts. While it is said that some data instances in the physionet set contained known noise sources such as dialogs which corrupt true heart sounds, trying to remove them via listening to recordings and manually editing out corresponding data intervals would entail a substantial amount of curation efforts. While automatic methods were proposed for the removal of dialogs and other artifacts from heart sounds, their reliability was somewhat short of serving to establish ground truth for sound signal portions [11]. Hence, sound intervals were not distinguished based on noise corruption criteria and raw recordings were used verbatim. Normality and abnormality are categorical distinctions. For the sake of computational feasibility, normal data label was assigned the numerical value -1 , and the abnormal ones $+1$. A subwindow of 7.5 seconds duration was extracted from each data instance. Some data lengths were shorter than the subwindow length. In this case, we arbitrarily replicated the available sound data portion and appended the replica to the end until the resulting stretched data exceeded or equaled the subwindow length. While this may introduce some artifacts of high frequency components in the resulting spectrum, which were not present in the original data, it expediently served to ensure signal length uniformity across dataset for feasibility of computations. Considering that only relatively small number of data had durations shorter than that of the subwindow and the acceptable level of performances levels achieved using the processed data, this replication tactic may well be justified.

2.1. Peak Detection

We first attempted to verify the reliability of peak detection methods using a detection tool availed by Physionet [12]. The official method tended to find peaks that are often absent from actual waveforms, possibly due to the assumption of the presence of four heart sound states of $s1$, systole, $s2$ and diastole in sequence and employment of the logistic regression Hidden markov models [2, 13]. Alternative methods discriminate peaks from non-peaks by focusing on local amplitude levels and do not necessarily assume global sound states. Other methods for peak detection were tested and found to have somewhat high rates of false positives and false negatives [14]. We developed a custom peak detection method to infer the occurrence of the S1 and S2 peaks. The nature of heart signal is such that the proportions of time sample points for S1 and S2 are far smaller than those of the non-S1 and non-S2 intervals (Fig. 1). This suggests the proportion information may be used towards enhancing the sensitivity and specificity of peak detection methods. Amplitude values were sorted by their absolute magnitudes and the level corresponding

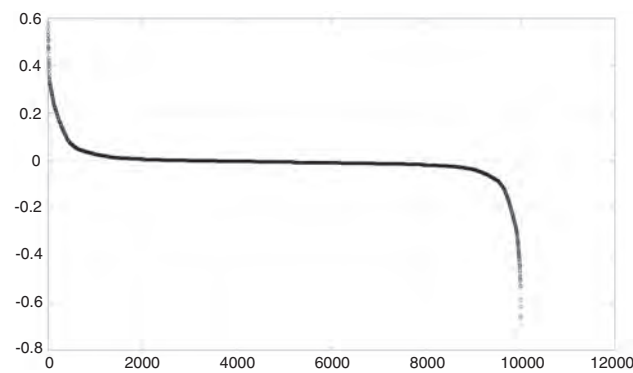


Fig. 1. Amplitude distribution of a heart sound.

to highest 2.5% of data points was used as threshold for discerning peaks from non-peaks (Fig. 1). Figure 2 illustrates a peak detection result obtained using our custom method.

2.2. Fourier Transform

Full length Fourier transform is known to be a venerable way to investigate the frequency components that constitute time domain signals. The results of Fourier transform respectively applied to normal and abnormal signals are shown in Figure 3. From manual inspection of Fourier transform results, differences between normal and abnormal data samples were less clear than those in time domain representation. In addition to the spectral peaks corresponding to the time domain S1 and S2 peaks, less dominant peaks were found irregularly over the spectra for normal and abnormal sounds. The subpeak locations were quite diverse in the respective normal and abnormal sound classes. Hence we did not pursue the use of full length fourier transform further in our study. Short time fourier transform is an effective way of expressing energy distribution of sounds in time-frequency domain. Time signals of 20,000 samples points were extracted from the middle of raw heart sounds in time domain. Single Fourier transform of 512 points was extracted from each window of size 256 with 128 sample point overlap (Fig. 4).

2.3. Amplitude Normalization

It is known that stethoscopic method often produces low level of sound amplitudes originating from the mechanical closure of minute body parts (heart valves). Also, specific locations of application of stethoscope and configuration of recording devices

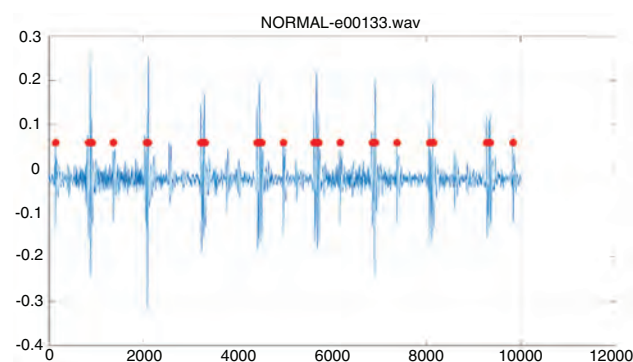


Fig. 2. Peaks detected by using custom method.

employed during the diagnostic session may negatively affect the amplitudes of heart sound signals. It may be conjectured that the absolute level of sound amplitude should not directly translate to basis for judging the normality of heart sound, considering the variability of recording protocols of participating medical agencies. Hence, for amplitude normalization scheme, we measured the highest absolute amplitude, amp_max , of the subwindow for each data and normalized the rest of samples in the subwindow by amp_max . Henceforth, the resulting sound data set is termed Amp_normal set. The PCG set of raw amplitude intact is termed Amp_intact set.

2.4. Deep Learning

For time domain data, signal points from the middle of raw data frame were extracted for each sound data, and were saved as bmp image. Each signal was downsampled by a factor of 5, and the signal portion of duration of 7.5 seconds was extracted to result in a subwindow of 3000 samples. CaffeNet model was used to train a classification model using the data. Its architecture is shown in Figure 5. While the data provided were labeled either normal or abnormal, the degree of abnormality varies rather gradually from mild, moderate to severe. Considering the inherent inhomogeneity and varying degree of pathological severity of data instances of the abnormal class, we used regression as the output layer of CaffeNet. By applying a threshold value to the output of regression network, a classifier can be obtained. For caffeNet training, $batch_size$ parameter value was set to 48. Other parameter values were as follows: initial learning rate = $3e-6$, step lr_policy , $gamma = 0.1$, $stepsize = 800$, $momentum = 0.9$ and $weight_decay = 0.0005$. Dropout rate was 0.50.

GoogLeNet model was also built, yet the performance difference between the two models was rather insignificant: in fact, CaffeNet tended to consistently produce performance figures slightly exceeding those of GoogLeNet. It may be hypothesized that relative simplicity of one dimensional time domain signal and possibly insufficient data used together with complex structure of deep learning network may result in overfitting such that a simpler deep learning model is preferred. Given the marginally higher performance, relative simplicity and less computational requirement of the CaffeNet, we used it throughout our study.

2.5. Ensemble

Individual deep learning networks can be combined in a range of ways to form a composite classifier that may be more effective from the constructive utilization of relative strengths of each classifier. Various ways of combinations are possible. We formed an ensemble classifier consisting of three distinct deep neural network classifiers in cascade (Fig. 6). The comprising classifiers were respectively trained on Amp_normal PCG set, amp_intact PCG set and STFT set. Alternatively, ensemble classifiers with component classifiers in parallel configurations were also built and tested (Fig. 7)—linear discriminant analysis (LDA) and support vector machine (SVM). LDA and SVM are similar in that both try to find separators that separate one set of data points from the other with presumed optimality criteria. For the present problem, normal data points constitute one set and the abnormal data points constitute another. LDA assumes normal distributions for each of data sets and finds a separator with maximal normalized distances from their respective distributions (Fig. 8).

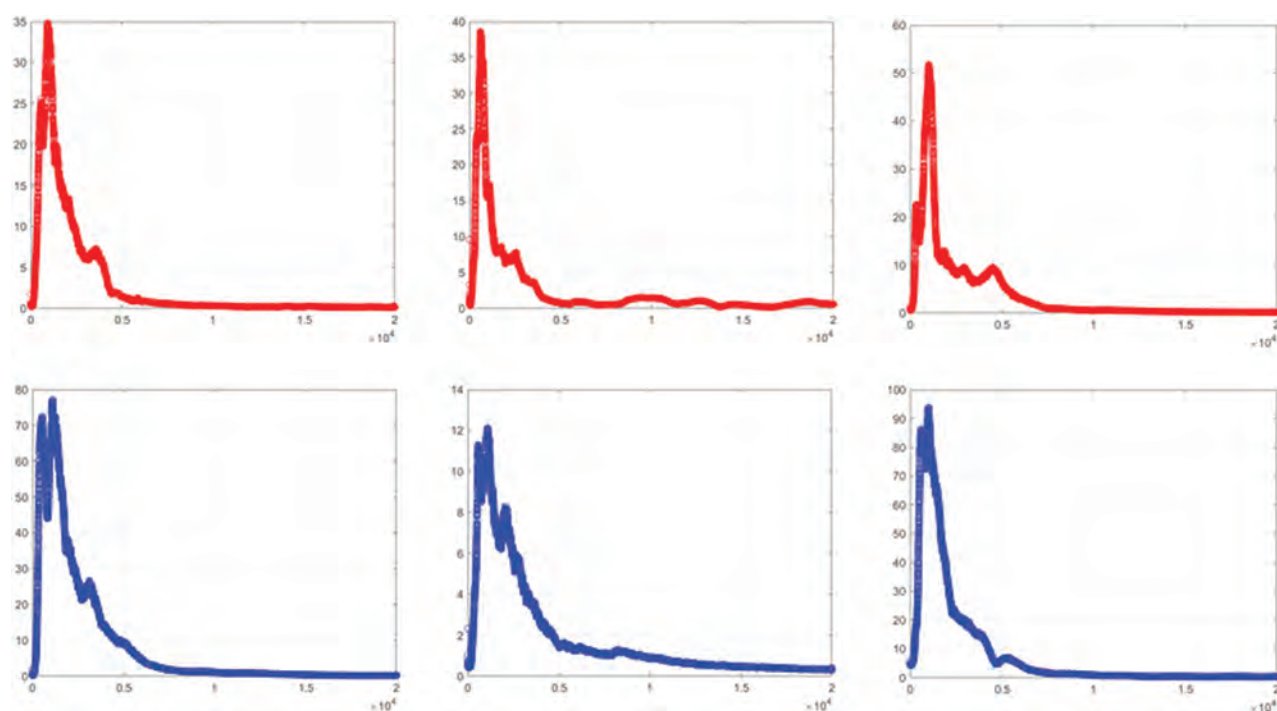


Fig. 3. Fourier transform results.

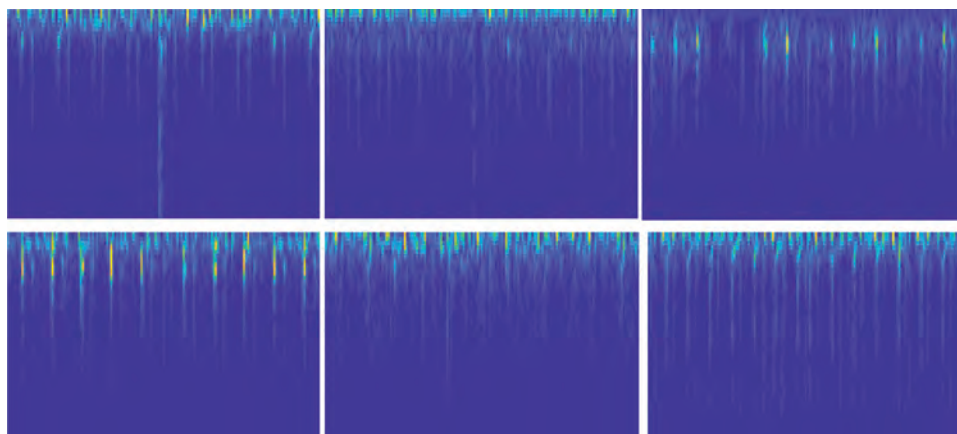


Fig. 4. Short time fourier transforms of heart sound signals.

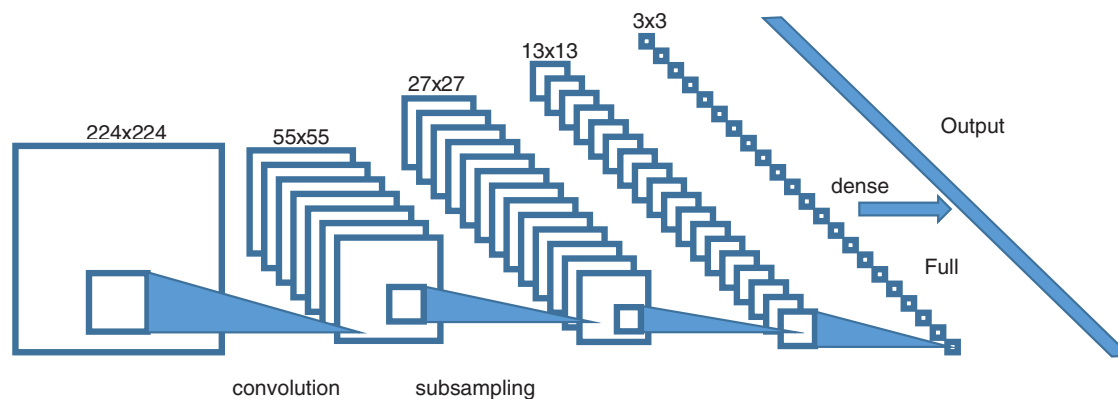


Fig. 5. CaffeNet architecture.

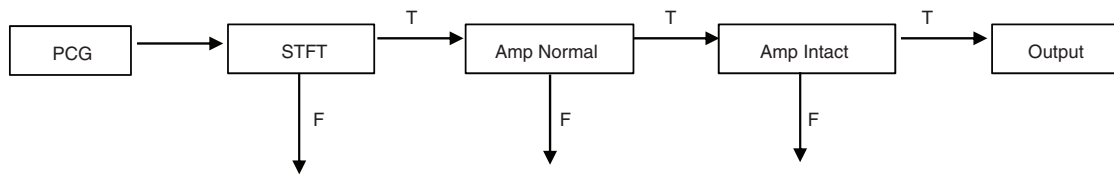


Fig. 6. A serial configuration of ensemble of classifiers based on time domain and frequency domain representations.

SVM finds separator with minimal sum of distance errors for mis-classified data points, without any assumptions on data distribution (Fig. 9). Often data points may not necessarily follow normal distribution, then the SVM often tend to outperform LDA in such circumstances.

2.6. Amplitude Normalization

It is known that stethoscopic method often produces low level of sound amplitudes originating from the mechanical closure of minute body parts (heart valves). Also, specific locations of application of stethoscope and the configuration of recording devices employed during the diagnostic session may negatively affect the amplitudes of heart sound signals. We conjectured that absolute level of sound amplitude should not directly translate to basis for judging the normality of heart sound, considering the variability of recording protocols of participating medical agencies. Hence, for amplitude normalization scheme, we measured the highest absolute amplitude, amp_max , of the subwindow for each data and normalized the rest of samples in the subwindow by amp_max . Henceforth, the resulting sound data set is termed Amp_normal set.

3. RESULTS

3.1. Peak Loc

When the peak locus alone was used as a representation of heart sound signal and Caffenet was trained with the loci image as input, the sensitivity was 0.8038, specificity was 0.4863 and overall accuracy was 0.4774. Specificity level was quite low to be considered for practical uses. It is known that abnormalities such as aortic ejection, opening snap and various kinds of systolic murmurs are mainly manifested in phonocardiogram as pathological, aberrant signals in the systole and diastole intervals. While the relative locations of peaks are quite important and vital for the overall integrity and functionality of heart, the result implies mere location of peaks is far short of reliably conveying the discriminative information relevant to heart normality/abnormality classification.

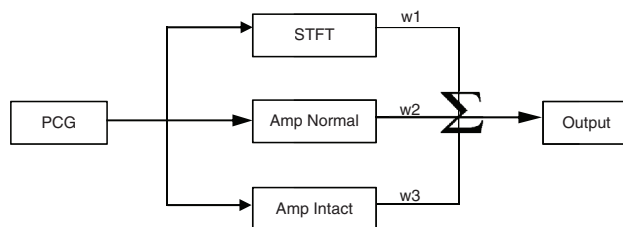


Fig. 7. A parallel configuration of ensemble classifiers based on time domain and frequency domain representations.

3.2. STFT

In addition, we obtained STFT representation of sound signals some examples of which are shown in Figure 4. The performance resulting from STFT representation with Caffenet was somewhat lower than that of plain time domain image representation (sensitivity = 0.7875, specificity = 0.7935 and accuracy = 0.7913).

3.3. Time Domain PCG

Utilizing regression Caffenet on time domain signal in image form alone produced 0.8860 precision and 0.9050 recall. Corresponding accuracy was 0.8825. There are ways to model time domain signal such as hidden markov model (HMM) and recurrent neural networks (RNN), etc. While sound signal is one dimensional, merely representing it as such may be suboptimal for achieving specific classification objectives. Considering sound signal essentially is continuous and digital sampling process typically produces a large number of samples, modeling the signal characteristics via discrete states of HMM or RNN may be hard or infeasible or the resulting model may incorporate many irrelevant modeling artifacts.

The structural peaks of S1 and S2 often showed substantial jitters across data or were often missing depending on the quality of data and on the normality/abnormality of corresponding subject. Plain two dimensional image representation of phonocardiogram (PCG) spanning multiple heartbeat cycles retains, in a single coherent form, the interrelations among $s1$, $s2$, systole and diastole as well as presence of murmurs from pathological causes. The modeling capacity and the expressiveness of two dimensional deep learning neural network can be effectively used to integrate such information to form a classifier for heart sound classification task. Considering model development efforts, the end to end manner of deep learning process facilitates to expediently form classification model, a great feature from practical perspective. A comprehensive table of performance values are given in Table I.

3.4. SVM

Overall, the SVM exhibited the highest performance level of 0.9012 accuracy, 0.9175 sensitivity and 0.8886 specificity. This was attained at the threshold value of -0.0690 . From Figure 10, the highest accuracy value is seen to coincide with the local maxima of classification accuracy versus threshold curve which somewhat resembles a Gaussian. Corresponding receiver operating characteristic (ROC) plot is shown in Figure 11. The AUC was 0.7218. It suggests some information relevant to classification missing either from the raw sound representation or STFT is present in the other, and the deep learning and SVM classification methods effectively utilizes such information towards the classification objective. On the other hand, cascade ensemble classifier produced somewhat higher error rates

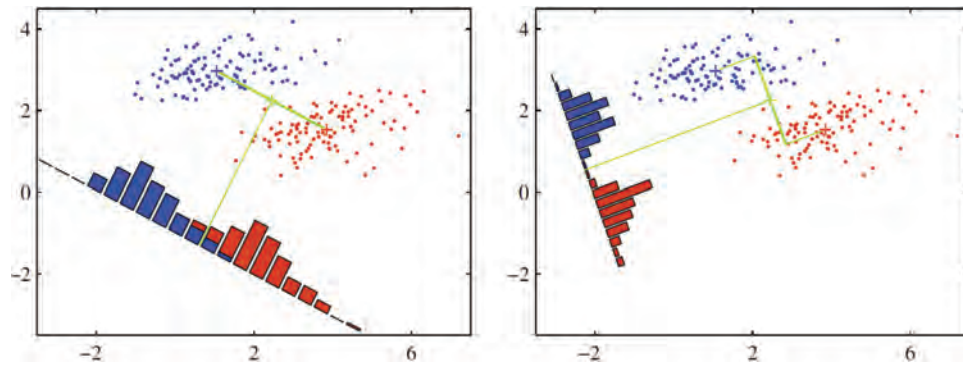


Fig. 8. Data classification by linear discriminant analysis.

than its constituent intact amplitude regression Caffenet alone (Table I).

With SVM, by adjusting threshold values, distinct models of different sensitivity and specificity values can be obtained.

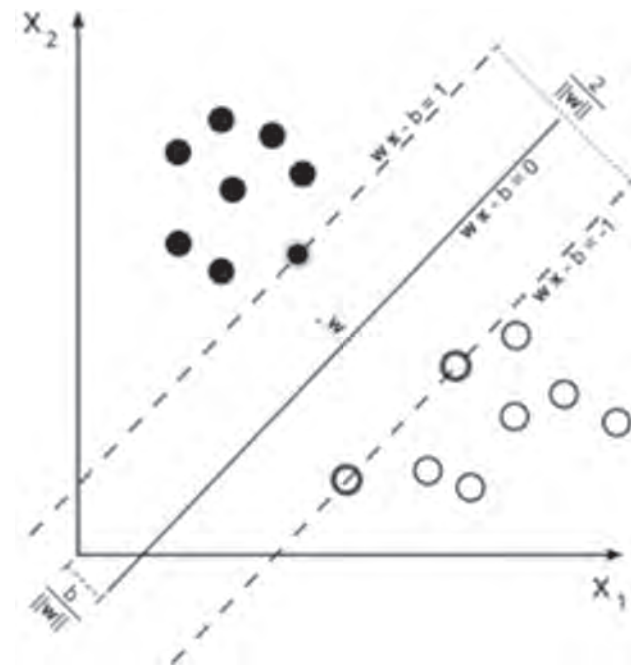


Fig. 9. Data classification by support vector machine.

Table I. Performances of different models.

	Time/freq	Sensitivity	Precision	Accuracy
Potes et al. [9]		0.9424	0.7781	0.8602
Caffenet-amp intact	Time	0.9050	0.8660	0.8825
Caffenet-amp normal	Time	0.8025	0.7887	0.7937
Googlenet-amp normal	Time	0.8566	0.7747	0.8038
Googlenet-amp intact	Time	0.8491	0.7785	0.8038
Caffenet	Peak locus	0.8038	0.4863	0.4774
Caffenet	STFT	0.7875	0.7935	0.7913
Ensemble-cascade	STFT + time	0.8975	0.8609	0.8763
Ensemble-parallel, LDA	STFT + time	0.9125	0.8881	0.8988
Ensemble-parallel, SVM*	STFT + time	0.9175	0.8886	0.9012
Ensemble-parallel, SVM*	STFT + time	0.9900	0.6226	0.7950
Ensemble-parallel, SVM*	STFT + time	0.9500	0.7917	0.8500

Note: * Same model with varying threshold values.

For threshold value of -0.5407 , the sensitivity of SVM model was 0.9900, specificity 0.6226 and accuracy 0.7950. With threshold value -0.2128 , respective performance values were 0.9500, 0.7917 and 0.8500. Since mis-diagnosing as normal a patient having real heart issues will have critical consequences from personal and ensuing medical cost concerns, such occasions would have to avoided which would then put relatively higher emphasis on sensitivity than specificity. The achieved high sensitivity levels with decent specificities imply proposed method will be able

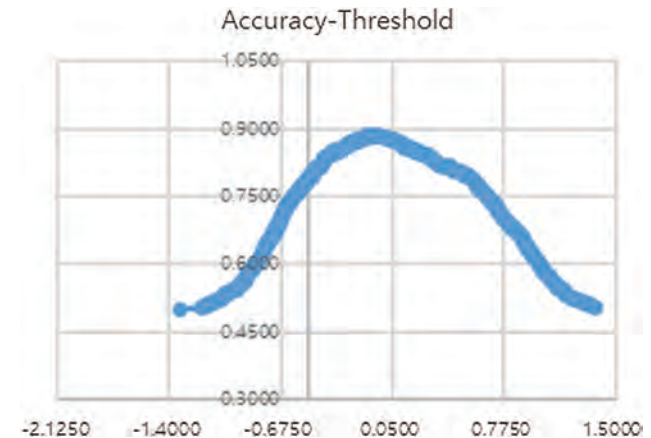


Fig. 10. Threshold-accuracy plot for determination of optimal threshold value in terms of accuracy criteria. The optimal threshold value was -0.069 for accuracy of 0.9012.

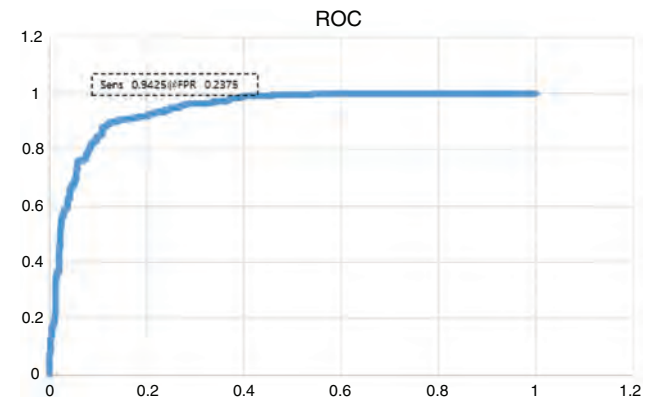


Fig. 11. Receiver operating characteristic (ROC) plot.

to meet such requirement. The performance figure at the highest accuracy level fares quite well with that of Potes et al. which ranked first in the Physionet 2016 heart sound classification challenge [9]. Even though the test set employed for the challenge incorporated two additional subsets that were not included in the publicly released sets and test figures obtained using the sets may be worse than the ones solely using available data, the results still suggest the potential of our composited deep learning-based method in modeling the difference between normal and abnormal heart sounds.

4. DISCUSSION

The adjustability of the threshold value suggests flexible application scenarios where specificity can be traded off for sensitivity depending on the requirement of the application and the nature and the availability of complementary diagnostic means afforded to medical staff. Competitive performance and the flexibility in threshold value setting suggest very promising uses of the proposed method as a first hand screening means in practical clinical environments. In many data instances, the S1 and S2 peaks were missing or hard to find. In this case, it is not meaningful to find peaks using peak detection methods which strongly tended to spot peaks regardless of their absences or presences. Often many kinds of heart murmurs resulting from heart anomalies span fairly wide frequency spectra.

Then the application of frequency selective filters to enhance certain bands may not be justifiable hence we did not apply such filters for the present study. Still representing over multiple separate spectra may be effective in detecting certain kinds of heart anomalies and may play a complementary and constructive role, which we may investigate in the future. Some sound data instances labeled abnormal had strong and clear S1 and S2 signals that were spaced in close to regular manners along time axis, yet featured emergent aberrant murmurs in either systole or diastole intervals.

It is noteworthy that the CaffeNet based model occasionally tended to classify such data instances as normals by assigning regression scores around the vicinity of -1 . A number of different approaches are conceivable to remedy this, including one involving a convoluted architecture of serial and parallel

processing branches to take effect depending on appropriately chosen factors including peak amplitudes. By segregating input data by criteria such as signal quality, peak presence, intensity and interval regularity, such misclassified data may be more accurately classified. We may also pursue such unattempted approaches in the future.

References and Notes

- Schelbert, E.B. and Messroghli, D.R., 2016. State of the art: Clinical applications of cardiac T1 mapping. *Radiology*, 278(3), pp.658–676.
- Liu, C., Springer, D., Li, Q., Moody, B., Juan, R.A., Chorro, F.J., Castells, F., Roig, J.M., Silva, I. and Johnson, A.E.W., 2016. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12), pp.2181–2213.
- Constant, J., 2003. *Essentials of Bedside Cardiology*, Springer.
- Loscalzo, J., Wiener, C.M., Bloomfield, G.T., Fauci, A.S., Braunwald, E., Kasper, D.L., Hauser, S.L. and Longo, D.L., 2008. *Harrison's Principles of Internal Medicine: Self-Assessment and Board Review*, McGraw-Hill Medical.
- Ranganathan, N., Sivaciyan, V. and Saksena, F.B., 2007. *The Art and Science of Cardiac Physical Examination*, Springer.
- Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems*.
- Sak, H., Senior, A. and Beaufays, F., 2014. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling, arXiv:1402.1128.
- Latif, S., 2018. Abnormal Heartbeat Detection Using Recurrent Neural Networks. *Computer Vision and Pattern Recognition*, arXiv:1801.08322.
- Potes, C., Parvaneh, S., Rahman, A. and Conroy, B., 2016. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Computing in Cardiology*, 43.
- Rubin, J., Abreu, R., Ganguli, A., Nelaturi, S., Matei, I. and Sricharan, K., 2017. Recognizing Abnormal Heart Sounds Using Deep Learning, arXiv:1707.04642.
- Paul A.S., Wan E.A. and Nelson A.T., 2006. Noise Reduction for Heart Sounds Using a Modified Minimum-Mean Squared Error Estimator with ECG Gating. *Conference Proceedings of IEEE Engineering Medical Biological Society*, 2006(1), pp.3385–90.
- Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov P.C., Mark R.G., Mietus J.E., Moody G.B., Peng, C.K. and Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), pp.E215–E220.
- Springer, D.B., Tarassenko, L. and Clifford, G.D., 2016. Logistic regression-HSMM-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4).
- Liu, C., Springer, D. and Clifford, G.D., 2017. Performance of an open-source heart sound segmentation algorithm on eight independent databases. *Physiological Measurements*, 38(8), pp.1730–1745.

Received: 21 May 2018. Revised/Accepted: 17 August 2019.