# Fundamental Heart Sound Classification using the Continuous Wavelet Transform and Convolutional Neural Networks

Andries Meintjes, Andrew Lowe, and Malcolm Legget

***Abstract*—Correct identification of the fundamental heart sounds is an important step in identifying the heart cycle stages. Heart valve pathologies can cause abnormal heart sounds or extra sounds, and an important distinguishing feature between different pathologies is the timing of these extra sounds in the heart cycle. In the design of an understandable heart sound analysis system, heart sound segmentation is an indispensable step. In this study classification of the fundamental heart sounds using continuous wavelet transform (CWT) scalograms and convolutional neural networks (CNN) is investigated. Classification between the first and second heart sound of scalograms produced by the Morse analytic wavelet was compared for CNN, support vector machine (SVM), and k-nearest neighbours (kNN) classifiers. Samples of the first and second heart sound were extracted from a publicly available dataset of normal and abnormal heart sound recordings, and magnitude scalograms were calculated for each sample. These scalograms were used to train and test CNNs. Classification using features extracted from a fully connected layer of the network was compared with linear binary pattern features. The CNN achieved an average classification accuracy of 86% when distinguishing between the first and second heart sound. Features extracted from the CNN and classified using a SVM achieved similar results (85.9%). Classification of the CNN features outperformed LBP features using both SVM and kNN classifiers. The results indicate that there is significant potential for the use of CWT and CNN in the analysis of heart sounds.**

## I. Introduction

The two fundamental heart sounds (FHS) being the first heart sound (S1) and the second heart sound (S2), respectively indicate the start and end of ventricular systole. The identification of these sounds, a process called heart sound segmentation, is the first step in the process of cardiac auscultation, or the analysis of heart sounds. Once these sounds have been identified, the states of systole and diastole within the heart cycle are revealed. Any extra sounds, such as murmurs, clicks, or snaps, which are present in the heart sound must be interpreted in the context of their position within these heart cycle states. The heart is a biological and thus highly variable system and the heart sounds may vary considerably even in a single individual, for example between states of calm and excitement, or performing movements that temporarily change the shape of the heart chambers or the position of the heart relative to the chest wall. The locations on the chest wall where the heart sounds are auscultated or recorded also influences the attributes of the FHS. Identification of the heart sounds is further complicated by the presence of disease conditions that may alter aspects of the heart sounds themselves or cause extra sounds. Finally, the environment in which the heart sounds are auscultated/recorded inherently contains noise sources and obstacles to effective heart sound analysis, from biological sources such as lung or gastric sounds, to artifacts produced by movement of the stethoscope/recording device and polluting environmental noises such as talking, noises of doors closing, people moving about, etc.

Methods of computer aided cardiac auscultation, or automatic heart sound analysis have become more and more applicable in recent years due to advancements in electronic stethoscopes, in particular the reduction in size and increase in processing power of microelectronics [1]. The current state of the art segmentation algorithm, used in the 2016 Physionet heart sound classification challenge [2], is based on work done by Springer et al. [3] and Schmidt et al. [4], [5]. This algorithm uses logistic regression and a hidden semi-Markov model to determine the locations of the fundamental heart sounds and the phases of the cardiac cycle in a heart sound recording. Simultaneously recorded ECG data are used to annotate the training set for the Markov model.

Sound event classification is a field of signal processing that attempts to identify the sources of sound events. A common approach used in sound event classification is to use time-frequency transforms to extract a time-frequency representation of the sound, after which features extracted from the transform are used to identify the most probable source of the sound. Short time Fourier transform (STFT) spectrogram based feature extraction along with feed-forward artificial neural networks (ANN) and k-nearest neighbor (kNN) classifiers were investigated in [6] to automatically identify urban sounds. Similarly in [7], environmental sounds were classified using a support vector machine (SVM) based on local binary pattern (LBP) features extracted from spectrogram transforms. Speech emotions were determined in a similar way in [8] using convolutional neural networks (CNN) as the classifier. Different time frequency representations of the fundamental heart sounds were investigated in [9]. The authors concluded that between the STFT, Wigner distribution, and continuous wavelet transform (CWT), the CWT gave the clearest representation of the time-frequency content in the heart sounds.

This study addresses classification of the FHS in recordings of normal and pathological heart sounds, made in noisy environments. It is possible to extract a large number of examples of the FHS from publicly available datasets, making deep machine learning methods (e.g. CNN) possible. The main contributions are generally an investigation into the usefulness of the CWT and CNN's in heart sound analysis

A. Meintjes and A. Lowe are with the AUT Institute of Biomedical Technologies, Auckland, NZ (phone: +6421 25 89 490; e-mail: ameintje@aut.ac.nz; alowe@aut.ac.nz).

M. Legget is with the Faculty of Medical and Health Sciences, University of Auckland, Auckland, NZ, (e-mail: malcolm.legget@auckland.ac.nz).

and more specifically a methodology of distinguishing between the first (S1) and second (S2) heart sounds using CWT decomposition and CNN features.

## II. Methodology

### A. Database and development environment

The data used in this study was retrieved from Physionet [10]. Specifically, the dataset from the LR-HSMM segmentation algorithm [3], released as an example dataset for the 2016 Physionet Challenge [2], was used. This dataset consists of 792 individual heart sound recordings sampled at 1000 Hz from 135 patients varying between 1 second and 35.5 seconds (a total of 9975 seconds) as well as annotations of R-peak and end-T-wave locations from electrocardiogram (ECG) recorded in parallel. Data were divided according to patient and classifiers were trained on a randomly selected 120 patients and tested on the remaining 15; this was repeated 9-fold. This ensured that all patients formed part of the testing set at least once. Although the number of examples per patient were not equal, the ability of the classifier to identify the FHS in unseen patients gives a more realistic measure of its performance.

The software environment used for this research was MATLAB R2018a using the Wavelet Toolbox, Neural Network Toolbox, and Statistics and Machine Learning toolbox. The CNNs were trained on a NVIDIA GTX 760 with 2GB of RAM.

### B. Preprocessing and fundamental heart sound extraction

All audio recordings were normalized by subtracting the mean and dividing by the standard deviation of the signal. [5] Samples were then filtered using a low-pass Butterworth filter with a 25 Hz cutoff to limit low frequency noise. The homomorphic envelope, calculated as the exponential of the natural logarithm of the Hilbert transform (1) of the signal [3], [11], was computed for each recording. The Hilbert transform is used to calculate a complex valued analytical signal, avoiding the singularity at log(0).

$$h(t) = x(t) * (\pi t)^{-1} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(u)}{t-u} du \qquad (1)$$

Examples of the first (S1) and second (S2) heart sound were extracted from the 792 samples in the development dataset. The average duration of S1 is 122ms (± 22ms) and of S2 is 92ms (± 22ms) [3], [12]. S1 was identified by finding the peak of the homomorphic envelope in a 122ms time window after the occurrence of the R-peak in the ECG. A 150 ms window centered at this peak was specified as S1. S2 was identified by finding the peak value in the envelope around the T-wave in the ECG. S2 was then specified as a 150ms time window centered at this peak. The duration of 150ms was chosen to ensure that the heart sounds were completely captured, based on the aforementioned durations of the heart sounds. A total of 11633 examples of S1 and 11574 examples of S2 were extracted from the database.

### C. The continuous wavelet transform

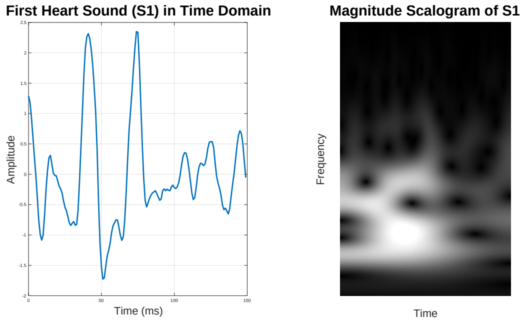
The CWT of a signal $s(t)$, can be defined as:

$$Z(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\psi^{*}\left(\frac{t-b}{a}\right) dt \qquad (2)$$

Where $s(t)$ is a finite energy signal, $\psi^{*}$ is the complex conjugate of the mother wavelet, and $a$ and $b$ are parameters that respectively control the scaling and translation of the wavelet. Large $a$ (scale) values expand the wavelet in time revealing low frequency information in the signal, while smaller scale values shrink the wavelet and reveal high frequencies present in the signal [13]. The continuous wavelet transform is calculated by varying the variables $a$ and $b$ continuously over the range of scales and length of the signal respectively.

Scalograms were created for all of the extracted heart sounds (Figure 1). A scalogram is a visual representation of the CWT of a signal, similar to a spectrogram created using a short time Fourier transform (STFT). The CWT provides superior time and frequency resolution to the STFT, effectively by allowing different size analysis windows at different frequencies. The scalograms of the signals clearly represent the frequencies present at different times in the signal and provide a visual representation that can be used to distinguish between the heart sounds. All CWT's were calculated using the Morse analytic wavelet with 48 voices per octave which resulted in 207 frequency windows between 23 and 423 Hz. Each scalogram was stored as a 207x150 pixel grayscale image.

*Figure 1:* An example of an extracted first heart sound (S1) and its corresponding time-frequency representation scalogram.



### D. Convolutional neural network

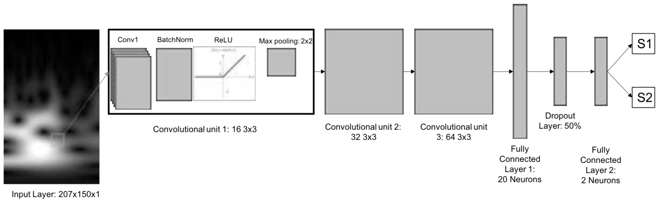
Convolution neural networks (CNN), also called deep convolutional neural networks (DCNN) [14], or ConvNets [15], [16], are machine learning classification tools that consists of one or more convolutional layers before a classification layer. CNN's have been found to be powerful image classifiers and have been used extensively in the research field of image classification since a deep convolutional neural network won the ImageNet Large Scale Visual Recognition (ILSVR) competition in 2012 [17].

Convolutional layers consist of spatial filters. During the forward pass, these filters are convolved across the width and height of the input (the raw image or output of the previous layer). Using supervised learning the weights of these convolutional layers are trained to extract features relevant to the classification problem. Thus CNNs do not require manual feature selection making them well suited to image classification problems in which the distinguishing features are not immediately obvious [15]. Other layers that are commonly used in CNNs are rectified learning units (ReLU) layers that act as non-linear activation functions, as well as pooling layers that downsample the input to reduce the number of parameters and also reduce overfitting [16]. Batch normalization layers are placed between convolutional and ReLU layers. These layers normalize the activation of each channel, reducing training time and sensitivity to network initialization [18]. The final layer of a CNN used for classification purposes is usually a layer of fully connected neurons that computes the class scores.

The architecture for the CNN used in this research is shown in Figure 2. In this figure the convolutional, batch normalization, ReLU, and pooling layers are visually combined into convolutional units to allow for clearer representation. Two fully connected layers of 20 neurons and 2 neurons respectively are used for classification. A dropout layer that performs regularization with a neuron retention rate of 50% is used in between these to reduce overfitting. A softmax function is used to compute class scores from the final fully connected layer.

A mini-batch size (number of samples in each training iteration) of 64 was used, and the networks were trained for 10 epochs (1 epoch is one full pass through the entire training set). Stochastic gradient decent with an initial learning rate of 0.01 and a piecewise drop of 0.2 every 2 epochs was used.

*Figure 2: Architecture of the convolutional neural network used in this study. Convolutional units 2 and 3 have the same contents as unit 1*

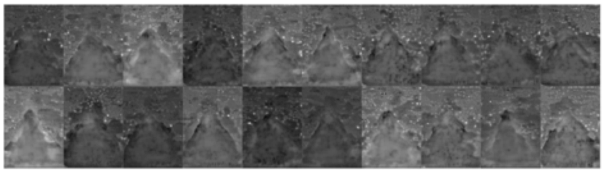

Conv1
BatchNorm
ReLU
Max pooling: 2x2
Convolutional unit 1: 16 3x3
Convolutional unit 2: 32 3x3
Convolutional unit 3: 64 3x3
Fully Connected Layer 1: 20 Neurons
Dropout Layer: 50%
Fully Connected Layer 2: 2 Neurons
S1
S2
Input Layer: 207x150x1

## *E. Feature extraction*

During the training process, the convolutional layers of a CNN start to activate or respond to parts of the input that correspond to distinguishable features between the classes. The initial convolutional layers act like edge detectors identifying very low-level features while the layers closer to the end of the network become sensitive to more complex/high-level features. The fully connected layers at the end of a CNN represent high level combinations of the features learned in the earlier layers of the network[16], [19].

The activations of the first fully connected layer of a trained CNN were extracted as features for FHS classification. This resulted in a feature of length 20, one value for each neuron in the fully connected layer. The fully connected layer is visualized in Figure 3.

*Figure 3: Visualization of a 20 neuron fully connected layer of a CNN trained to classify S1 and S2. Each rectangle in the 2×10 matrix represents a high-level combination of CNN features.*



For comparison purposes local binary pattern (LBP) features [20] were extracted from the scalograms. A radius parameter of 1 was used which resulted in 59 features per image.

## *F. Fundamental heart sound classification*

The features extracted from the CWT scalograms using CNN's were classified using the final fully connected neural network and softmax function of the CNN itself, support vector machine (SVM) classifiers, and k-Nearest Neighbours (kNN) classifiers. For a comparison between classification based on CNN features and other established features, the LBP extracted features were classified using SVM and kNN classifiers.

Classifiers were assessed on their accuracy, sensitivity, specificity, Area-Under-Curve (AUC) of the Receiver-Operator Characteristic curve, and net reclassification improvement (NRI) measure. All measures are reported with S1 taken as the positive class. NRI is a measure of the percentage of samples classified correctly by one model compared to another [21].

# III. RESULTS

The results of the FHS classification averaged over all 9 folds are shown in Table 1 and Table 2. NRI scores reported in Table 1 represent the reclassification score of the classifier on the left compared to the classifier on the right of the first column.

The average training time for the CNN's was 14 (±3) minutes. CNN feature extraction took an average of 62 (± 5) seconds. The LBP features took an average of 8 (± 30 sec) minutes to extract. Training times for the SVM classifiers were 26 (±9) and 8 (±1) s for CNN and LBP features respectively. The kNN training times were 31 (± 11) ms and 38 (±25) ms for CNN and LBP features. All reported times are for the entire training set (average 20628 ± 712 samples).

TABLE 1: NET RECLASSIFICATION IMPROVEMENT BETWEEN CLASSIFIERS AND FEATURES

| | *NRI* |
|---:|---:|
| *CNN vs SVM* | 0.23% |
| *CNN vs kNN* | 7.61% |
| *SVM vs LBP SVM* | 10.59% |
| *kNN vs LBP kNN* | 19.10% |

TABLE 2: PERFORMANCE MEASURES FOR THE CLASSIFICATION OF S1 AND S2 SCALOGRAMS.

| | **Sensitivity**<br>**mean**<br>**(std)** | **Specificity** | **AUC** | **ACC** |
|---|---|---|---|---|
| *CNN* | 87.4<br>(3.8) | 86.7<br>(5.0) | 93.8<br>(2.5) | 86.0<br>(3.4) |
| *SVM* | 87.65<br>(3.2) | 86.1<br>(4.9) | 92.7<br>(2.5) | 85.9<br>(3.3) |
| *kNN* | 82.9<br>(4.3) | 81.5<br>(5.0) | 77.4<br>(13.3) | 82.2<br>(2.9) |
| *LBP SVM* | 81.6<br>(6.0) | 80.9<br>(3.3) | 88.3<br>(3.7) | 80.5<br>(3.9) |
| *LBP kNN* | 73.5<br>(6.4) | 72.0<br>(5.1) | 72.7<br>(3.2) | 72.8<br>(3.2) |

## IV. DISCUSSION

In this research, the ability of convolutional neural network features to distinguish between time-frequency representations of the first and second heart sounds was investigated. The complete deep CNN, including the final fully connected layer had the best overall performance (accuracy = 86.0%) but did not perform statistically significantly better than the SVM classifier trained with CNN features (accuracy = 85.9%, p = 0.391). FHS classification using CNN features outperformed classification using LBP features on both the classifiers that were investigated (SVM: 85.9% vs 80.5%, p=0.0009; kNN: 82.2% vs 72.75%, p= 0.00001). Convolutional neural network training time was significantly higher than any of the other classifiers, although once trained the feature extraction using CNN was significantly faster than LBP feature extraction (62 sec vs 8 min).

The inherent variability of heart sound recordings makes the limited size of the data set a hindrance to the design of a generalizable algorithm. The data used in this study were recorded at a single facility and by a limited number of physicians which may not be sufficiently representative of the recording practices of the wider medical population. A larger dataset from varied sources could potentially produce more generalizable networks.

From the results it is clear that the FHS are significantly distinguishable based on their isolated time-frequency representations. The samples of S1 and S2 used in this study contained various sources of noise as well as pathological sounds. Despite this, the classifiers and in particular CNNs were able to correctly label the sounds in a majority of cases. A network able to accurately distinguish between the FHS could be applied in a heart sound segmentation algorithm to classify sound events in a heart sound under analysis as S1 or S2. Such networks could also be used to provide an independent measure of how successfully a heart sound segmentation algorithm (e.g. [3]) has identified the heart sounds, by extracting the heart sounds that the segmentation algorithm has identified, classifying them as either S1/S2 and comparing this to the results produced by the segmentation algorithm. Another possible application of such networks is as part of an object detection algorithm that recognizes the FHS in a scalogram of a heart sound recording. More generally the results indicate the usefulness of the continuous wavelet transform and convolutional neural networks in the analysis of heart sounds.

## V. CONCLUSION

Accurate identification of the fundamental heart sounds is the most important step in heart sound segmentation and an integral step in overall heart sound analysis. This study has shown that features identified by convolutional neural networks are able to distinguish between the fundamental heart sounds using continuous wavelet transform scalograms of the isolated sound events. This study provides promising results for the use of CWT and CNN in the development of heart sound analysis systems. Future work includes integrating FHS CNNs into heart sound segmentation and heart sound analysis and description algorithms.

## REFERENCES

[1] S. Leng, R. S. Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong, "The electronic stethoscope," *Biomed. Eng. OnLine*, vol. 14, Jul. 2015.

[2] "Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016." [Online]. Available: https://www.physionet.org/challenge/2016/. [Accessed: 22-Jan-2018].

[3] D. Springer, L. Tarassenko, and G. Clifford, "Logistic Regression-HSMM-based Heart Sound Segmentation," *IEEE Trans. Biomed. Eng.*, pp. 1–1, 2015.

[4] S. E. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent Hidden-Markov Model," 2008, pp. 345–348.

[5] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiol. Meas.*, vol. 31, no. 4, pp. 513–529, Apr. 2010.

[6] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," *Inf. Sci.*, vol. 243, pp. 57–74, Sep. 2013.

[7] K. Z. Thwe and N. War, "Environmental sound classification based on time-frequency representation," in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2017, pp. 251–255.

[8] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, 2017, pp. 1–5.

[9] S. m. Debbal and F. Bereksi-Reguig, "Time-frequency analysis of the first and the second heartbeat sounds," *Appl. Math. Comput.*, vol. 184, pp. 1041–1052, Jan. 2007.

[10] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[11] I. Rezek and S. Roberts, *Envelope Extraction via Complex Homomorphic Filtering*. 1998.

[12] A. G. Tilkian and M. B. Conover, *Understanding Heart Sounds and Murmurs: With an Introduction to Lung Sounds*. New York, NY, USA: Elsevier Health Sciences, 2001.

[13] John Sadowsky, "The continuous wavelet transform: a tool for signal investigation and understanding.," *John Hopkins APL Tech. Dig.*, vol. 15, no. 4, pp. 306–318, 1994.

[14] A. Teramoto, T. Tsukamoto, Y. Kiriyama, and H. Fujita, "Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks," *BioMed Res. Int.*, pp. 1–6, Aug. 2017.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[16] "CS231n Convolutional Neural Networks for Visual Recognition." [Online]. Available: http://cs231n.github.io/convolutional-networks/. [Accessed: 22-Jan-2018].

[17] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[18] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *ArXiv150203167 Cs*, Feb. 2015.

[19] D. Garcia-Gasulla *et al.*, "On the Behavior of Convolutional Nets for Feature Extraction," Mar. 2017.

[20] "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell. Pattern Anal. Mach. Intell. IEEE Trans. IEEE Trans Pattern Anal Mach Intell*, no. 7, p. 971, 2002.

[21] E. S. Jewell, M. D. Maile, M. Engoren, and M. Elliott, "Net Reclassification Improvement," *Anesth. Analg.*, vol. 122, no. 3, pp. 818–824, Mar. 2016.