MACHINE LEARNING AND WEARABLE DEVICES FOR PHONOCARDIOGRAM-BASED DIAGNOSIS

Shaima Abdelmageed, Prof. Mohammed Elmusrati

School of Technology and Innovation, University of Vasa, Vasa, Finland

ABSTRACT

The heart sound signal, Phonocardiogram (PCG) is difficult to interpret even for experienced cardiologists. Interpretation are very subjective depending on the hearing ability of the physician. mHealth has been the adopted approach towards simplifying that and getting quick diagnosis using mobile devices. However, it has been challenging due to the required high quality of data, high computation load, and high-power consumption. The aim of this paper is to diagnose the heart condition based on Phonocardiogram analysis using Machine Learning techniques assuming limited processing power to be encapsulated later in a wearable device. The cardiovascular system is modelled in a transfer function to provide PCG signal recording as it would be recorded at the wrist. The signal is, then, decomposed using filter bank and the analysed using discriminant function. The results showed that PCG with a 19 dB Signal-to-Noise-Ratio can lead to 97.33% successful diagnosis.The same decomposed signal is then analysed using pattern recognition neural network, and the classification was 100% successful with 83.3% trust level.

KEYWORDS

Analysis, Classification, data quality, diagnosis, filter banks, mHealth, PCG, SNR, transfer function, Wavelet Transform, wearable

1. INTRODUCTION

This paper presents the analysis of the heart's acoustic signal at the wrist. To find this signal, a transfer function is proposed to represent the impact of the travelled distance from the heart to the wrist on the characteristics of the PCG. This signal is expected to have low quality and low signal-to-noise ratio. To reduce the required computation load and power consumption and speed up the processing time the signal is downsampled by 100 sample/s before it is decomposed using filter bank into four subbands. Each subband is described using two features; mean and covariance. The system is trained using 300 cases to diagnose the heart condition against six hypothetical diseases. The classification is based on the discriminant function of the unclassified signal. The most probable diagnosis is found by maximising the discriminant function and in other words, minimising the Mahalanobis distance [1]. For the experiment, nonstationary noise is used to simulate nonstationary environment such as the chaotic environment of accidents. The same signal is decomposed using 4-level Filter Bank and then analysed using pattern recognition Neural Networks (NN) to validate the concept.

2. LITERATURE REVIEW

There is a lot of research in the area of diagnosing heart conditions, especially in the last ten years. For instance, in 2016, a team of two attempted building heart sound monitor using

wearable wrist sensor [2]. Although they focused on the hardware, the mathematical model of the heart acoustic system was based on the analytical approach discussed here. The wrist sensor was designed to detect heart pulses by measuring pressure changes to calculate the heartbeats. As previously discussed in this research, the sound wave changes with the change of blood pressure, which is inevitable when the sound wave propagates through the blood vessels (following the blood flow), this applies to pulse wave as well. In [2], they worked on finding the inverse function for the pulse wave between two locations along the same artery; chest and wrist and used that to estimate the recorded pulse wave at the chest from the one recorded at the wrist. Figure 1 shows the model that was presented in [2], it represents the travel of the pulse wave sound in blood vessels. The job is now about inversing the effect of the blocks in this model, once the pulse wave was captured at the chest and the wrist, the two were compared to find the relation between them. The delay correlates with the distance travelled from the heart to the wrist; they relied on the pulse peaks in time domain to estimate the delay using Short-Time Fourier Transform (STFT). They then trained a two-layer feed-forward backpropagation neural networks to find the inverse attenuation function. The main task here was to validate the recorded pulse wave at the wrist and consequently validate the proposed hardware.
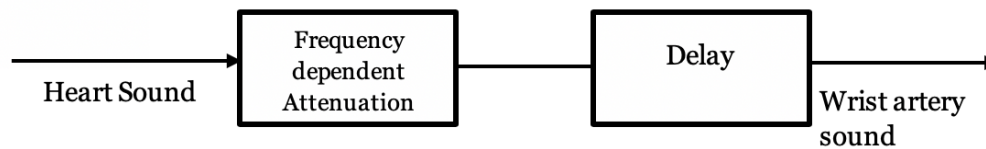

Figure 1. Travel model of pulse wave sounds in blood vessels

3. OBTAINING PCG AT THE WRIST

This is achieved by modelling the heart-wrist acoustic wave propagation system using SIMULINK MATLAB, see Figure 2 for the model. A healthy heart acoustic signal with sampling date of 44100 sample/s is used as a reference, it is downsampled by 100 sample/second to reduce the computation load and speed up the processing time. It should be noted that down sampling acts as a filter that removes the high frequencies, which works well in the case of low frequencies like the one at hand. The new resultant signal has slower sample rate than the original signal. This could affect the accuracy of the experiment negatively, since the down sampling of the signal could lose some of the disease's indications that are held in high frequencies. All information and indications held beyond 200 Hz are lost, the down sampled signal shows only 200 frequency components. Not to mention that it would limit the number of levels that can be in the Filter Bank [3], since some sub bands would show only noise. However, the purpose of this paper is to diagnose the phonocardiogram signal (PCG) using machine learning with limited computation load and processing power (mobile devices), down sampling serves this purpose.After which, a random noise is added to account for measurement noise, the total signal is used as an input to the model. The resultant signal at output of the model is a distorted version of the input, Figure 3 shows the original signal after downsampling and the resultant signal after going through the propagation model to simulate what the signal would be at the wrist.
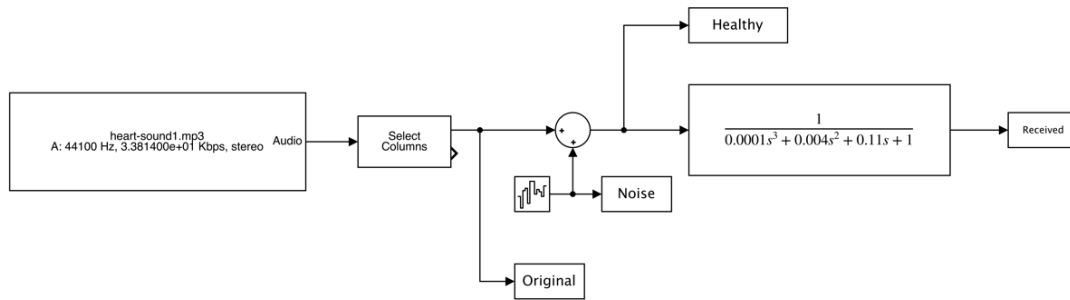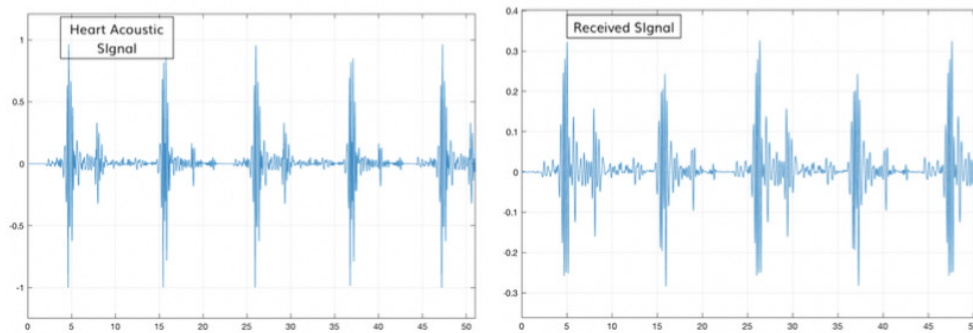
Figure 2. Heart-wrist acoustic propagation model

Figure 3. Original and received signal at the wrist

3.1. GENERATING THE HYPOTHESES

The experiment starts by defining five hypotheses, each assumed to represent a heart disease, these hypotheses are used as reference classes for the diagnostic system. This is done by introducing a unique transfer function that alters the healthy heart acoustic signal in Figure 3 to form a hypothetical disease, the resultant is then sent through the heart-wrist model (acoustic propagation transfer function) to be received as hypothesis x. The hypotheses are scientifically valid because, from an engineering point of view, an unhealthy heart is a heart that produces faulty PCG. When the cardiovascular system has a problem, the acoustic signal of the heart should reflect that in time/frequency space. This appears as corruption in the heart sound; such change in the acoustic signals might not be audible or sensed by human ears (not even experienced cardiologist). However, it could be sensed and potentially classified with sensitive sensors and proper machine learning algorithms. A heart disease could be modelled as a linear corruption of signal of the healthy sound. This corruption has been performed by applying the healthy heart sound signal to different linear filters with different characteristics. There is no medical basis for selecting these filters during the simulation. However, this could be another research for modelling different heart problems with linear/nonlinear transfer functions.The diagnostic system's job is to identify this noise and consequently conclude the most probable condition. This is perhaps one of the limitations of this experiment, it did not involve medical researcher nor ethical approval to use real heart acoustic signals. The research was carried out by an engineer and had to rely on simulated signals based on one real acoustic signal of a healthy heart.

4.   THE EXPERIMENT

The experiment (parametric estimation) is to apply the proposed solution to classify 300 cases. All cases are generated from the hypotheses discussed in 3.1, below are the steps of the experiment. Figure 4 shows the block diagram of the proposed system.
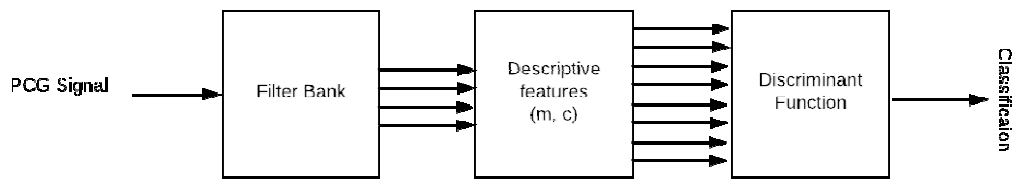
Figure 4. PCG-based diagnostic system

No restoration technique was used in this work, because the distortion level in the received signal was reasonable. Not to mention that any restoration technique would have added to the computation load and processing time, which goes against the goal of this paper.

4.1. DECOMPOSING WITH FILTER BANK

The signal is decomposed using Wavelet Transform and specifically Filter Banks [2,3] with different number of levels for every cycle of this experiment. The training and test sets were decomposed into a number of bands equivalent to 2x Filter Levels. For each band, the mean and covariance were calculated, so that each case is described by a number of features equivalent to 2x the number of bands. These features are concatenated in a 1 x the number of features matrix named "descriptive matrix". The number of levels was selected using trial and error as the method of optimisation. Starting the 1 level (2 bands), the experiment trials and compare results from 2 levels (4 bands), 3 levels (8 bands), 4 levels (16 bands), 5 levels (32 bands), and finally 6 levels (64 bands). Increasing the levels beyond 6 did not add any value to the classification nor did it improve the classification result, for that this experiment was stopped at 6 levels.

4.2. TRAINING THE SYSTEM

Training the system in this approach is about using the training set to construct the discriminant function. By calculating the mean of the descriptive matrix to get a 1x128 vector and calculate the covariance of the descriptive matrix, which is a square matrix of 128x128. These values are used to construct the Discriminant Function (DF) [1] that is given by Equation 1

$$(1) \qquad DF=-0.5 \log(\det(s))- 0.5 (x-m)'*inverse(s)*(x-m)$$

Where s is the covariance matrix, m is the mean vector, and x is the training case. It certainly helped to add a confirmation step here that tests the DF using the training set, by simply calculating the DF for the training set and maximising the result, in order to confirm the validity of the training. This is a simple test; because if the training is valid the classification must be 100% correct. During the confirmation step, it became evident that the determinant of the covariance matrix is zero in many cases, which made the first component of the DF function (-infinity). For that, the discriminant function equation was rewritten as follow

$$(2) \qquad DF=-0.5 (x-m)' * inverse(s) * (x-m)$$

This formula was used when the filter levels reached 4 (16 bands, 32 features).

4.3. TESTING THE SYSTEM

To test the system, the discriminant function calculated in step (2) is used to classify the test set. The classification was 97.33% successful using 8 features (filter bank with 2 levels), with only 4 false classifications. With 128 features (6 levels), the classification was 79.33% successful. Total

of 31 cases out of the 150 cases were falsely classified, while the rest 119 were correctly classified.

4.4. DISCUSSION

After repeating the above steps for every all Filter Levels from 1 to 6, it was concluded that the best possible configuration for this scenario using this approach is to use a Filter Bank with 2 levels. This decomposes the signal into total of 4 bands and allows the signal to be described by 8 features. This is an interesting result because it defies the proposition behind this approach, that is more features will improve the success rate of the classification. Despite defying the proposition of the approach, this result is eye-opening. Decomposing signals is considered insightful, because it provides more descriptive details about the original signal. When studying the correlation coefficients between the cases with different decomposition levels, it was evident that the first band of every hypothesis correlate with the others, these bands are from a 6-level filter bank (64 bands and 128 features). This is visible in Figure 5-a, where the scatter plot shows that apart from some outliners, there is a positive linear correlation with moderate-strength across the first band of every hypotheses for majority of the values within these bands.As opposed to what can be seen in Figure 5-b, where the scatter plot shows little to no correlation across the first band of every hypotheses from a 2-level filter bank (4 bands and 8 features).
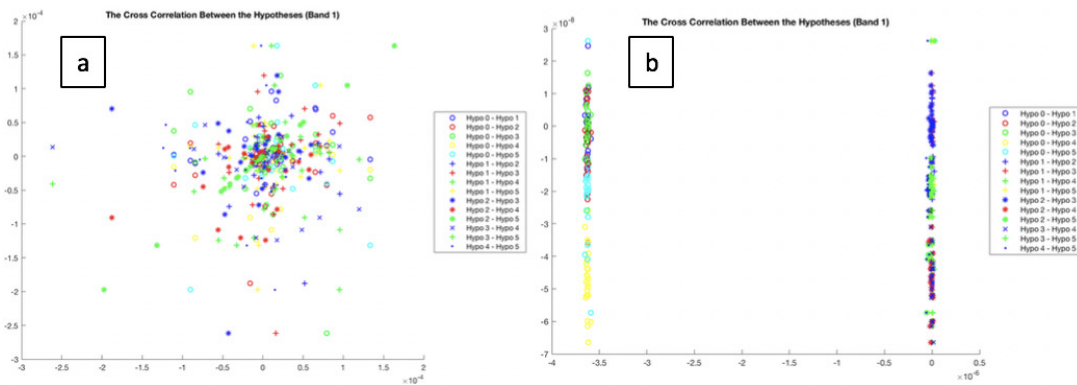

Figure 5. Correlation between subbands in different decomposition levels.

From this discussion, it can be concluded that smaller bands tend to have stronger correlation across the hypotheses, which causes the classification to be more error prone. Such bands result from decomposing the signal using filter banks with large number of levels. This could have been caused by the down sampling, since it removed information held in high frequencies and left some bands with just noise. However, the trade-off was worthy, because the down sampling that was performed at the beginning of the experiment reduced the required computational load and processing power. Not to mention that it served the purpose of this paper; limited energy and processing power (the use of mobile devices). Similar result had been noted in image texture recognition that filter banks with smaller number of levels performed better than larger ones [4].

5.   USING NEURAL NETWORKS

The same decomposed signal using Wavelet Transform (Filter banks) is analysed using Patten Recognition Neural Network. This was done to validate the concept and explore the possibility of improving the performance. Figure 6 shows the diagnostic system using NN.
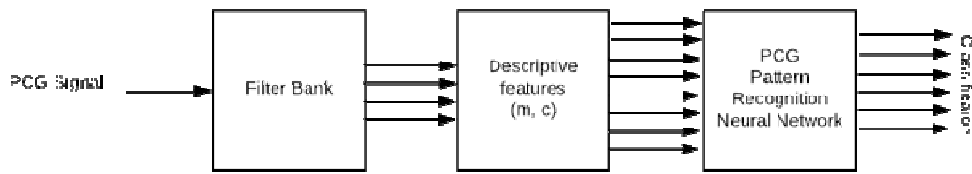
Figure 6. PCG-based diagnostic system (NN)

5.1. THE NEURAL NETWORK

The neural network is a two-layer feedforward network with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer. This simple network had one hidden layer with 10 neurons, Figure 7 shows the network diagram.
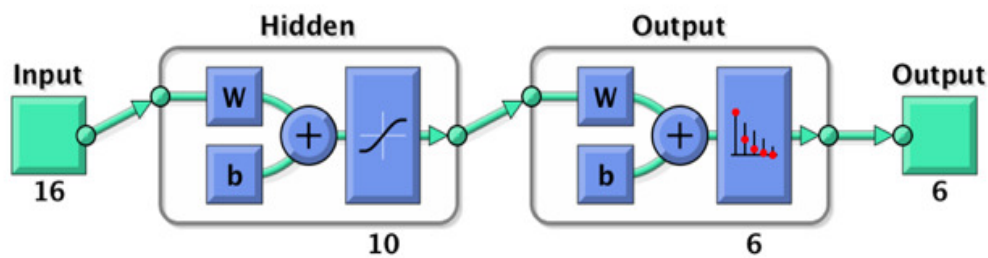

Figure 7. Neural Patten Recognition Network Diagram

There are many applications of Neural Networks in the area of Tele-cardiology [5,6,7,8].Moreover, more specifically, many applications used the duo of Filter Banks and Neural Networks for feature extraction and classification, respectively [9, 10, 11].

5.2. TRAINING THE SYSTEM

The data set had 150 samples with 16 elements for the input, which was generated using 4-level filter bank. Similarly, it had 150 samples for the target with 6 elements. each element represented a class (Hypo 0, Hypo 1, Hypo 2, Hypo 3, Hypo 4, Hypo 5), and each target vector showed '1' for the correct class and '0' elsewhere. The data set was divided into 70% for training, 15% for validation and 15% for testing. Consequently, the network output has 6 elements. For training, the network used Scaled Conjugate Gradient training algorithm. The training was successful and so was the testing and validation as can be seen in Figure 8.
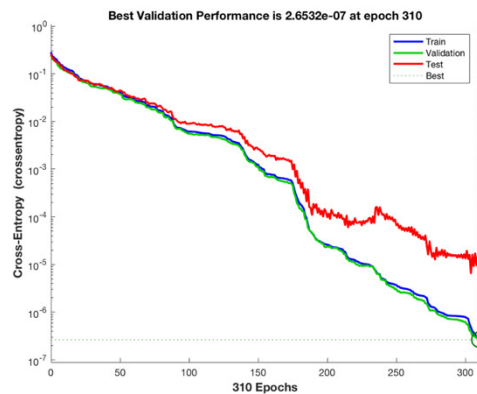

Figure 8. Performance plot

As can be seen in the confusion plot in Figure 9, 25 cases of every class were recognised, which represents all the dataset that was used for training, testing and validation, with zero% errors and zero confusion. That's when the training was concluded and there was no need for further testing. Following this, the network was used to classify a new dataset of 150 samples. The classification was 100% successful, the network was able to recognise the features of every sample with high accuracy; where the output showed 0.99999* for the correct class and <0.00001* for all other classes. The difference between the values of the output vector is very large, which proves low confusion and high performance.
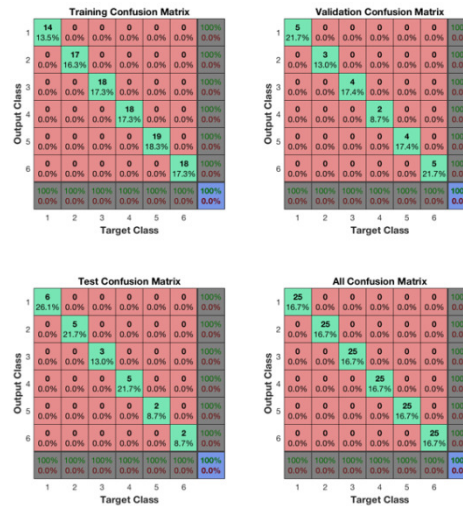

Figure 9. Confusion plot

5.3. TRUST LEVEL

Trust level is a proposed measure for how much can the classification result be trusted; the term is proposed by the author. Trust level should give context for the classification success. It is presented here using the formula in Equation 1.

$$Trust\ level = (Maxima – second\ maxima) – Confusion\ Index \qquad (1)$$

Where Confusion Index is the probability of each Hypothesis included in the training phase, which is 0.167 and from the resultant classification, the Maxima – second maxima= 1. Therefore, the trust level on the classification provided by this Neural Network is $1 - 0.167 = 0.833$, that is 83.3%. This is quite satisfactory; however, it could be improved by increasing the number of Hypotheses. The below discussion explains this further.

5.4. DISCUSSION

There are six classes that have equal presence in the studied sample. Therefore, the probability that a given case is either of these classes is 0.17 and since it is equal, it has been ignored. But let's consider a sample where the classes do not have equal presence, in other words, imbalanced sample.For example; assuming the following imbalanced sample was used for training

$$P(H_0|\hat{S}) = 0.1$$
$$P(H_1|\hat{S}) = 0.5$$
$$P(H_2|\hat{S}) = 0.1$$
$$P(H_3|\hat{S}) = 0.2$$

$$P(H_4|\hat{S}) = 0$$
$$P(H_5|\hat{S}) = 0.1$$

The training process would have been more complicated, since it could suffer from local optima, where it keeps classifying cases to the class with the highest presence (highest probability). Not to mention that the accuracy of the classification would become a secondary measure of the network performance, since the cases that have been classified as Class (e.g. H_1) are 50% correct. In fact, such sample would be more realistic than the balanced sample that was used in the experiment, since heart diseases, and diseases in general, have different probabilities. These probabilities are impacted by the gender of the patient, their age group, their life style, and for some diseases their origin and country of residence. To improve the performance of the classification of imbalanced sample such problem, some of a combination of the following techniques can be used;

- Collect more data from the underrepresented classes to reach a balanced sample. This will simplify the problem but might not always be possible. And in other times, it might not be desired, in case of limited resources.

- Simply duplicate training samples to match up the number of samples per class in the training dataset. However, this will increase the training time without adding any new information. This solution can have better impact if the copies are augmented and modified to add new information, for instance a distorted version of the same class.

- Remove training sample from the overrepresented classes to match up the number of samples per class in the training dataset. This will reduce the training time but will worsen the network performance by removing useful information for the training.

- Train for *sensitivity* and *specificity*. *Sensitivity* is the probability of classifying a case as H_1, while it is in fact H_1, which measures the accuracy of detecting a Hypothesis. While *Specificity* is the probability of classifying a case as H_1, while it is in fact H_2, which measures the accuracy of detecting the absence of a Hypothesis. These two attributes describe the accuracy of the classification as a whole. For optimal performance, they should both be equally high (~1) due to their equal importance. However, this might be different in some application. For example, skin cancer detection system is better off with high sensitivity and low specificity. The two attributes can be traded-off depending on the application.

- For every class, specify a weight that is inversely proportional to the probability of the class. For example, H_1 should have a weight of 0.5, while H_3 should have a weight of 0.8. These weights represent the contribution of the class to the loss function. A practical implementation of this was presented in 2015 and it is called "Keras Implementation" [12].

Another solution would be to introduce a new hypothesis "none of the known hypotheses". This gives room to define an acceptable probability of error, where any classification that is lower than a certain minima would lead to "none of the known hypotheses". This method increases the credibility and integrity of the system compared to the previous one, however, the fewer the hypotheses the more probable the neutral diagnosis is concluded. Moreover, the higher the acceptable probability, the more accurate the system is.In conclusion, the number of the learnt hypotheses has a huge impact on the accuracy level of the diagnostic system. The more knowledgeable the diagnostic system, the more accurate the diagnosis will be, which is achieved by training ANN on larger number of hypotheses.

6. CONCLUSION

For a chaotic environment, like the one represented with nonstationary noise, the proposed approach provided satisfactory results with ~97% success rate when using low-quality PCG signal was decomposed using 2-level Filter Bank. Thesuccess rate increased to 100% when pattern recognition neural network was used for classifying the PCG signal decomposed using 4-level Filter Bank. Describing the signal using16featuresinstead of 8 features using the parametric estimation proposed in the experiment (refer to section 4). Moreover, the Neural Network-based classification had a trust level equal to 83.3% which correlates with the number of hypotheses included in the classification. The PCG signal had low-quality of 19dB in this case as well, which makes it possible to be encapsulated into wearable device due to the low computational load and processing time. There is room for future work in the same area, perhaps using real PCG recordings to build databases of heart conditions and expand the learning.

REFERENCES

[1]   Alpaydin, E. (2016). Introduction to Machine Learning. 3rd ed. ed. Cambridge: The MIT Press.

[2]   Shi, W.Y. and Chiao, J.-. (2016). Neural Network Based Real-Time Heart Sound Monitor using a Wireless Wearable Wrist Sensor. IEEE Available from: http://dx.doi.org/10.1109/DCAS.2016.7791150

[3]   Mertins, A. (1999). Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications. Chichester [u.a.]: Wiley.

[4]   Randen, T. (1997). Filter and Filter Bank Design for Image Texture Recognition. Norwegian University of Science and Technology Stavanger College.

[5]   Mutijarsa, K., Ichwan, M.andUtami, D.B. (Oct 2016). Heart Rate Prediction Based on Cycling Cadence using Feedforward Neural Network. IEEE Available from: https://ieeexplore.ieee.org/document/7863026 Available from: http://dx.doi.org/ 10.1109/IC3INA.2016.7863026.

[6]   O.V. Melnik, Yu.A. Chelebaevaand S.V. Chelebaev. (2017). Real-Time Heart Rate Parameter Analysis Based on Artificial Neural Networks. 6th Mediterranean Conference On Embedded Computing. 6th Mediterranean Conference on Embedded Computing 11-15 JUNE 2017. IEEE.

[7]   Rastgar-Jazi, M. and Fernando, X. (Jul 2017).Detection of Heart Abnormalities Via Artificial Neural Network: An Application of Self Learning Algorithms. IEEE Available from: http://dx.doi.org/10.1109/IHTC.2017.8058202

[8]   Ali, S.Q. and Hossen, A. (2018). Different Neural Networks Approaches for Identification of Obstructive Sleep Apnea. IEEE.

[9]   Wang, Z., He, Z. and Chen, J. (1999). Filter Banks and Neural Network-Based Feature Extraction and Automatic Classification of Electrogastrogram. Annals of Biomedical Engineering, vol. 27, no. 1, pp. 88-95. Available from: http://dx.doi.org/10.1114/1.151

[10] Kadambe, S., Murray, R. and Boudreaux-Bartels, G.F. (1999). Wavelet Transform-Based QRS Complex

[11] Liang-Yu Shyu, Ying-Hsuan Wu and Hu, W. (2004). Using Wavelet Transform and Fuzzy Neural Network for VPC Detection from the Holter ECG. IEEE Transactions on Biomedical Engineering, vol. 51, no. 7, pp. 1269-1273. Available from: http://dx.doi.org/10.1109/TBME.2004.824131

[12] Chollet, F. and et el. (2015).Keras Documentation.

AUTHORS

Shaima Tajalsir Abdelmageed is a Doctoral student at the University of Vasa, Finland. She has been a researcher in the field of mHealth since 2011.

Prof. Mohammed Elmusrati is full professor and head of communications and systems engineering group at the University of Vasa. He is also a visiting professor at Aalto University. Prof. Mohammed is also supervising Shaima's Ph.D. research.