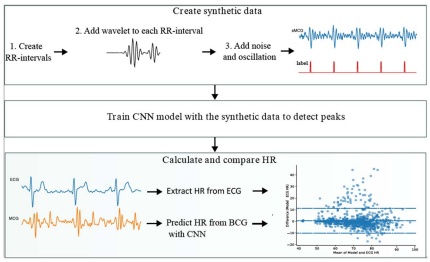
Mechanical sensors

Generating Synthetic Mechanocardiograms for Machine Learning-Based Peak Detection

Jonas Sandelin, Ismail Elnaggar, Olli Lahdenoja, Matti Kaisti, and Tero Koivisto
Digital Health Technology Group, University of Turku, 20500 Turku, Finland

Manuscript received 13 July 2024; accepted 2 August 2024. Date of publication 14 August 2024; date of current version 26 September 2024.

Abstract—Acquiring labeled data for machine learning algorithms in healthcare is expensive due to the laborious expert annotation and privacy concerns. This challenge is further complicated in the case of mechanocardiogram (MCG) data, which are characterized by high interpersonal and intrapersonal complexity, compounded further by sensor variability. In this letter, we introduce an innovative method for generating synthetic MCG signals to address the scarcity of labeled data necessary for training machine learning models in healthcare. Our approach involves generating RR-intervals, adding wavelets, and incorporating noise to create realistic synthetic MCG signals. These synthetic signals were used to train a convolutional neural network for peak detection in real MCG data. Our key contributions include developing a detailed methodology for realistic synthetic MCG signal generation, reducing the mean absolute error in peak detection by 4.88 beats per minute using synthetic data, enhancing the training of machine learning models, creating a new peak detection method, and addressing data scarcity in biomedical signal processing. These contributions emphasize the methodological innovations and the significance of our results, underscoring the potential impact of synthetic data in improving healthcare diagnostics.



Index Terms—Mechanical sensors, electrocardiogram (ECG), machine learning, mechanocardiogram (MCG), neural network, synthetic.

I. INTRODUCTION

In recent years, the use of deep neural networks in the field of healthcare has significantly increased [1]. However, these solutions demand a large amount of labeled data, a resource that is both challenging and time-consuming to gather [2]. This challenge is particularly pronounced in healthcare data, where expertise is required for accurate labeling and interpretation, and privacy issues further complicates the situation [3]. The problem is even more significant when it comes to the newer methods in the field, such as mechanocardiograms (MCGs). MCG refers to all mechanical cardiac monitoring sensors, including seismocardiography (SCG), ballistocardiography (BCG), and gyrocardiography [4], [5]. MCG signals describe the mechanical vibrations and movements of the heart [6]. One benefit of these sensors, compared with the gold standard electrocardiogram (ECG), is that they are much easier to obtain in home environments. Most people with a smartphone can record MCG signals, highlighting the accessibility of these biosignals. However, the variability in sensors used for data collection contributes to significant signal variance. Currently MCGs are employed in various clinical and research settings to screen heart conditions, monitor cardiac health, and study the mechanical properties of heart function [4], [5], [6], [7], [8], [9], [10], [11].

The morphological components of typical MCG signals consist of two distinct heart sounds, S1 and S2, which are shown in Fig. 1. Understanding the characteristics of S1 and S2 in the MCG signal is pivotal for interpreting the mechanical behavior of the heart, forming the basis for effective peak detection and further cardiovascular assessment. One of the challenges with MCG signals is peak detection. Due to the nature of the signal, it can easily be distorted by motion artifacts and noise [12].

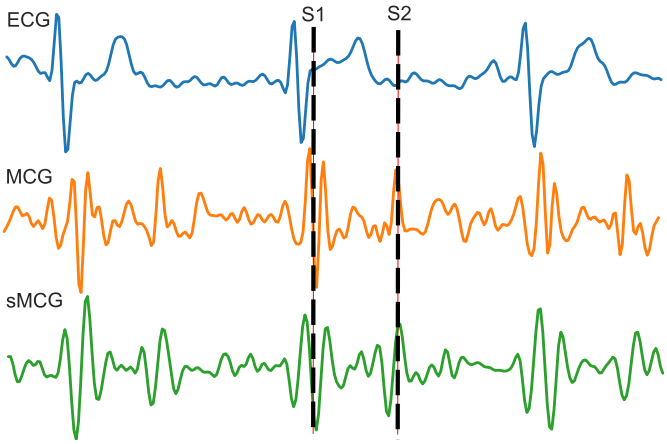


Fig. 1. Real ECG and MCG signals and a synthetic MCG signal created by using the RR-intervals derived from the ECG.

Since one of the challenges in MCG research is the scarcity of sufficient data, emulating sensors and creating synthetic data could be useful. The number of open datasets in MCG is minimal compared to the number of open ECG and photoplethysmography (PPG) datasets, for example. For MCG, there are fewer than ten open datasets, with the largest containing only 100 subjects [6], [9], [10], [11], [13], [14], while there are multiple open datasets for PPG and ECG with over 10 000 subjects [15], [16].

Since obtaining annotated datasets can be difficult, it can limit the development of robust models for clinical applications. Synthetic data

Corresponding author: Jonas Sandelin (e-mail: jojusan@utu.fi).
Associate Editor: M. K. Shukla.
Digital Object Identifier 10.1109/LSENS.2024.3443526

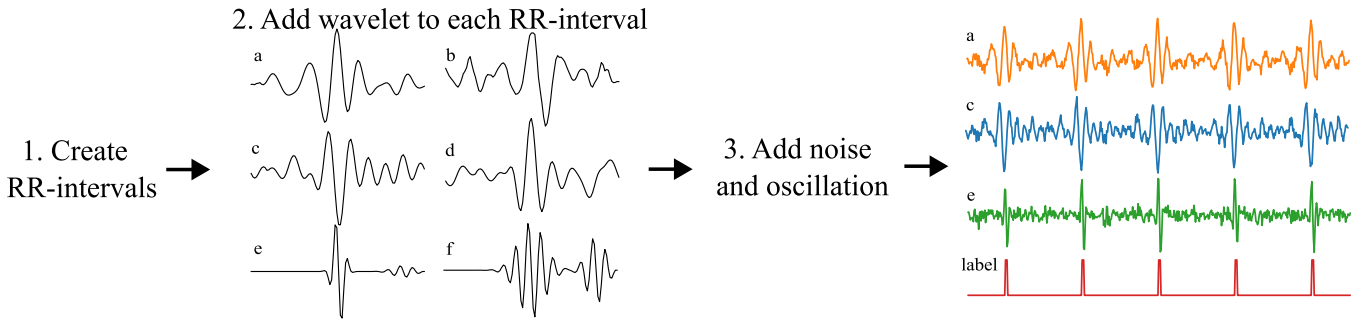


Fig. 2. Synthetic signal generation involves creating RR-intervals, selecting wavelets derived from real signals (a–d) or Gaussian sinusoids (e and f) and adding noise and oscillation. Wavelets e and f represent the minimum and maximum values for Gaussian sinusoid in our study. This process then results in MCG signals with corresponding labels. Example signals use wavelets a, c, and e.

generation emerges as a promising avenue to overcome these challenges by offering an unlimited supply of diverse MCG signals with precise annotations. It is also more cost-effective and time-efficient compared to the labor-intensive process of collecting and annotating real MCG data. Ethical and privacy issues are also reduced when generating the signals mathematically. Synthetic data also open up novel research opportunities by allowing the exploration of scenarios and conditions that may be difficult or impractical to study with real data alone. This capability can lead to new insights and innovations in cardiovascular health monitoring and diagnostics.

However, due to the high complexity of these mechanical signals, accurately simulating them can be difficult. MCG signals can vary from person to person due to weight, health, and other morphological factors [6], [9]. Moreover, individual morphological changes can manifest within the same recording or subject. This is particularly evident when sensor locations are altered [8].

Some previous studies have been published on creating synthetic MCGs with different methods [2], [12], [17]. These methods include, for example, a generator phantom, which replicates SCG signals physically with a speaker alongside synthetic ECG [12]. Second approach trains a generative adversarial network with real SCG heartbeats and then uses the model to generate synthetic SCG heartbeats [17]. Third approach is similar to the second, since it uses a transformer-based neural network trained with real SCG heartbeats to generate synthetic beats [2]. For other biomedical signals, such as ECG and PPG, synthetic data have already been proven to be a useful tool [18], [19]. These two studies used more of a algorithmic approach in order to generate synthetic signals, where they generate the waveforms for the signals based on mathematical functions. Our approach is a combination of these, combining algorithmic approach with the real life heartbeats.

In this letter, we introduce a novel approach that emulates real-life sensor characteristics to generate synthetic MCG signals. Subsequently, we validate the effectiveness of this method by training a convolutional neural network (CNN) model for peak detection with the generated data and test the model with real sensor data.

II. METHODS

A. Synthetic BCG Signal Generation

The BCG signal generator is divided into three main parts: RR-interval generation, beat generation, and noise pollution.

1. RR-intervals are generated based on a randomly determined heart rate (HR) between 50 to 130 beats per minute. The HR can vary 100 ms up or down during the length of the signal.
2. A wavelet is added to each RR-interval. This was done in two different ways in our algorithm. 1) We take an existing wavelet from a real signal and modify it randomly for each beat. We extracted four different wavelets from real signals, seen in Fig. 2(a)–(d). The wavelets are then modified using a Gaussian-modulated sinusoid [20]

$$e^{-at^2+j\cdot 2\pi f_c t}.$$

The length of the sinusoid is randomized between 15 to 40 ms, the amplitude from 0 to 1, and the frequency between 6 to 16 Hz. This enhances the realism and variability of the synthetic MCG signals. 2) We generate the beats using only Gaussian sinusoids. In this approach, the beats are generated entirely synthetically. Fig. 2(e) and (f), representing the minimum and maximum versions of these Gaussian sinusoids. The length (25–60 ms), frequency (6–16 Hz), and derivative (+/-) of the sinusoid are all randomized. In addition, an S2 is added as another Gaussian sinusoid to the signal 300–500 ms after the first sinusoid (S1). The amplitude of the S2 wavelet is randomized between 0.1 to 0.7 times the amplitude of S1, adding further variability to the synthetic MCG signals.
3. Finally, noise is added to the generated signals. First, we add white noise to the signal. This noise is created by a 1-D inverse discrete Fourier Transform [1] with the lower threshold randomized between 0.1 to 2 Hz and the upper limit between 5 to 100 Hz. After this, an oscillation signal is added to the signal. We observed that some of the recorded MCG signals had oscillations, and adding them to the synthetic signals improved the results and enhanced the resemblance to real MCG signals.

B. Peak Detection on Real Signals

1) Datasets: To evaluate our peak detection system, we utilized four distinct datasets, each containing MCG signals and an ECG reference. All signals were resampled to 100 Hz and filtered with a Butterworth filter to maintain consistency. ECG signals were filtered within the range of 0.5–20 Hz, while the MCG signals were filtered within the range of 0.5–10 Hz. All of the signals were recorded in a supine position.

The first dataset (I) includes 20 healthy subjects, collected in University of Turku by the authors using a ring device with six channels of seismo- and gyrocardiography signals. The ring was placed on the index finger of each subject. The average recording length is 5.5 min, and the sampling frequency is 100 Hz. For each subject, the best channel was manually selected by the authors by identifying the signal with the strongest heartbeats.

The second dataset (II) is an open dataset [9] that includes 20 healthy subjects. Each recording consists of an ECG and one channel of SCG

signal. Each subject has three recordings, with one used as a test recording and the other two included in the training set. The average length for the test data is 5 min, while the training data average 50 min. The signals were originally sampled at 5 kHz.

The third dataset (III) is another open dataset, containing six-axis seismo- and gyrocardiographic data from 29 healthy subjects [6]. The average recording length is 9 min, and the signals were originally sampled at 800 Hz. The best channel was chosen using a method similar to that used by the original authors, which involves calculating the highest peak-to-peak amplitude divided by the absolute median deviation.

The fourth dataset (IV) is an open dataset that includes 40 ballistocardiogram signals recorded using a bed-based sensor [10]. These recordings include eight channels of BCG data. The average recording length is 7 min, with the signals originally sampled at 1 kHz. The best BCG channel was selected using the same method as in dataset III, as it worked well.

2) Peak Detection: From all the datasets, R-peaks from the reference ECGs were detected using a modified Pan–Tompkins algorithm [21]. From these peaks, HR was calculated and used as the reference.

The HR for the MCG signals was calculated by detecting peaks with the CNN model. All signals were segmented into 4-s long segments that overlap with a step size of 1. After the segmentation, the segments were fed to the CNN model, which then iterates through the signal. The model predicts the peaks for each segment, from which we can calculate the mean prediction array for the whole signal. From this array, we can then detect the local maxima with SciPy's find_peaks function. The detected peaks were adjusted to find the local maxima within the original MCG signals to ensure that the peak is in the correct place. This was done by looking for the maxima within a small window next to the predicted peak. With these local maxima, we can then calculate the HR and compare it to the reference ECG.

C. Neural Network and Training

1) Convolutional Neural Network: A CNN model was trained to detect peaks from 4-s-long signals with a sampling rate of 100, resulting in 400 samples per signal. The model comprises seven convolution layers with exponential linear unit activation and a kernel size of 5. The dilation rate of each layer increases exponentially from 1 to 64. The final layer employs sigmoid activation for the 400-sample output, serving as the prediction array. The model utilizes the Adam optimizer and binary cross-entropy loss function, resulting in a model with 3281 parameters. All of the values and layers were determined through experimentation to balance size and effectiveness.

2) Training: Six models were trained using identical structures but different datasets. Each model was trained with a batch size of 2048 and 150 epochs. The first four models utilized synthetic signals with varying wavelet styles. Model 1 used waveforms from real signals, Model 2 used Gaussian waveforms without S2, Model 3 used Gaussian waveforms with S2, and Model 4 used a combination of all variations. Real data from database II were used for Models 5 and 6. Model 5 incorporated every other real signal along with 150 000 synthetic signals, following the same variation as Model 4, while Model 6 exclusively used real data.

III. RESULTS

The results were produced by comparing the HR derived from the ECG reference with the HR derived from our CNN model using the MCG signals. HR was calculated for every 30-s segment. Fig. 3 shows the Bland–Altman plot for all the databases. Most of the HR

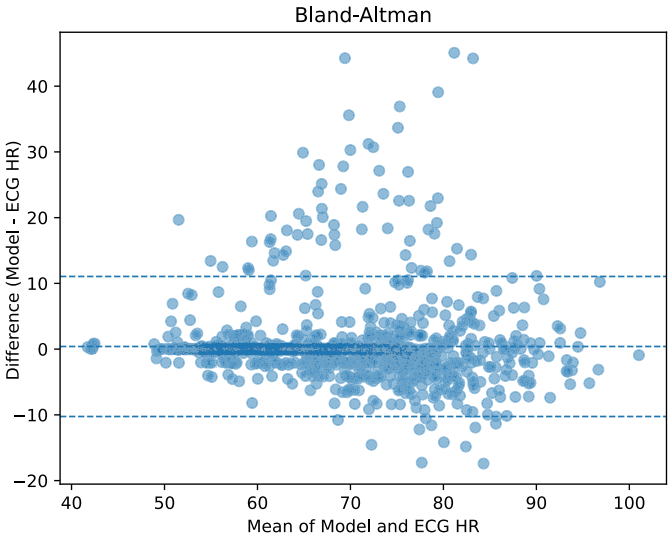


Fig. 3. Bland–Altman plot for all databases using Model 4.

Table 1. MAE of HR compared from ECG to BCG for all of the databases with different CNN models and state of the art reference

#	CNN model What data were used	Database *I*	*II*	*III*	*IV*
1	Real waveforms	4.66	3.21	6.19	7.62
2	Gaussian waveform, no S2	4.90	2.48	2.64	2.95
3	Gaussian waveform with S2	4.75	1.78	3.37	4.75
4	All of the above	**2.29**	**1.12**	**2.72**	**2.30**
5	Synthetic and real signals	4.86	2.37	3.09	3.48
6	Only real signals	4.59	6.21	9.25	7.92
7	**Autocorrelation**	**2.18**	**2.18**	**3.98**	**1.46**
4	Precision	86.11%	93.56%	73.92%	28.42%
4	Sensitivity	94.01%	93.22%	90.92%	26.99%

The best results for the synthetic model, and then the 'reference method' of AC is bolded.

are within the acceptable range, but some differ significantly from the reference HR.

In Table 1, comprehensive results across all databases are presented. The primary metric used for evaluation was the mean absolute error (MAE), chosen for its straightforward interpretation and ability to measure the average errors between the correct and predicted HR. The first four results present the synthetically produced signals with different variations, while Models 5 and 6 include real signals in the training set. Result 7 is a reference result, where the HR was calculated using an autocorrelation function. Autocorrelation has been found to be an effective method to estimate HR for healthy MCG signals [22], [23], [24].

Notably, employing any single variation in isolation did not yield optimal results. However, combining all data generation approaches into one model yielded the best results. It is worth emphasizing that certain databases exhibited varying responses to specific data variations, further underscoring the nuanced relationship between data characteristics and performance. It seems that adding variability and different variations increases the authenticity of the synthetic signals. The results also indicate that the model trained with real data does not perform as well as Model 4, which used a combination of synthetic data.

It is also important to note that the results calculated HR for short segments, meaning that missing one or predicting one extra peak can significantly affect the result. With a longer window, the result might be numerically better but would not accurately reflect the detection method.

Precision and sensitivity in Table 1 are calculated by comparing R-peaks from the ECG and the predicted peak from the MCG using Model 4. If these peaks are within 150 ms of each other, it is considered a true positive. If not, it is considered a false positive, and if there is no prediction for a peak at all, it is considered a false negative. The results are reasonably good; for Database III, it seems that we are predicting a lot of false positives. Comparing the result to that of the original study [6], it seems to be a bit worse but in line with their result. We, however, do seem to have more sensitive system than them. For Database IV, however, it seems that the peaks are not detected within 150 ms of the R-peak. When investigating the signals further, it becomes evident that the gap between R-peak and MCG peak is much longer. When doing the same test with a 400 ms, precision is 94.4% and sensitivity is 92.6%. This could mean that the peaks are detected correctly, but the delay is longer than with other MCG signals.

IV. CONCLUSION

The results of the study were generally good and demonstrated the potential of synthetic signals. We were able to produce realistic signals that can be used to train a peak detection CNN model. This peak detection method can successfully detect peaks from real MCG signals, indicating that the synthetic data closely resemble real data. We were able to see better model performance when augmenting our training data with synthetic signals. The performance increased especially when adding complexity and variability to the training data. Gaussian waveforms, in particular, increased the performance. High S2 especially seemed to be a challenge for some of the models, but adding big S2 into the training data increases performance.

A notable limitation of this letter is the small amount of test data, which included only healthy subjects. Consequently, the model's performance on subjects with heart diseases remains uncertain. Addressing this limitation requires expanding the test data to include a diverse range of subjects with varying health conditions. In addition, potential biases introduced by synthetic data generation need to be considered, as they may impact the model's generalizability and accuracy.

The results show that the peak detection system and the CNN model could be improved. Other models were tested during the study, but this model was chosen for its simplicity, computational speed, and effectiveness. Long short-term memory (LSTM) models also performed well but required significantly more training time. Future research could explore more complex models or solutions to fully leverage the synthetic data generator developed in this study. Additional directions could include improving the synthetic data generation process and exploring a more diverse range of synthetic data, including different arrhythmias. Testing the model with a completely new database could further validate its effectiveness and robustness. This would help in assessing the model's generalizability and performance across various conditions and populations.

ACKNOWLEDGMENT

This work was supported by the ITEA project called RM4HEALTH, Business Finland under Grant 8139/31/2022.

REFERENCES

[1] I. Zion, S. Ozuomba, and P. Asuquo, "An overview of neural network architectures for healthcare," in *Proc. Int. Conf. Math. Comput. Eng. Comput. Sci.*, 2020. pp. 1–8.
[2] M. Nikbakht et al., "Synthetic seismocardiogram generation using a transformer-based neural network," *J. Amer. Med. Informat. Assoc.*, vol. 30, no. 7, pp. 1266–1273, 2023.
[3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinf.*, vol. 19, no. 6, pp. 1236–1246, 2018.
[4] S. Sieciński, P. S. Kostka, and E. J. Tkacz, "Gyrocardiography: A review of the definition, history, waveform description, and applications," *Sensors*, vol. 20, no. 22, 2020, Art. no. 6675.
[5] R. Casanella et al., "Cardiac mechanical signals," in *Cardiovascular Computing—Methodologies and Clinical Applications*. Berlin, Germany: Springer, 2019, pp. 63–79.
[6] M. Kaisti et al., "Stand-alone heartbeat detection in multidimensional mechanocardiograms," *IEEE Sensors J.*, vol. 19, no. 1, pp. 234–242, Jan. 2019.
[7] F. Marc, "Feasibility of monitoring congestive heart failure with seismocardiography: A literature review," 2023. [Online]. Available: https://studenttheses.uu.nl/handle/20.500.12932/44205
[8] P. Suresh, N. Narayanan, C. V. Pranav, and V. Vijayaraghavan, "End-to-end deep learning for reliable cardiac activity monitoring using seismocardiograms," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl.*, 2020, pp. 1369–1375.
[9] M. A. García-González, A. Argelagós-Palau, M. Fernández-Chimeno, and J. Ramos-Castro, "A comparison of heartbeat detectors for the seismocardiogram," in *Proc. IEEE Comput. Cardiol.*, Sep. 2013, pp. 461–464.
[10] C. Carlson et al., "Bed-based ballistocardiography: Dataset and ability to track cardiovascular parameters," *Sensors*, vol. 21, no. 1, 2020, Art. no. 156.
[11] K. Munck, K. Sørensen, J. Struijk, and S. Schmidt, "Multichannel seismocardiography: An imaging modality for investigating heart vibrations," *Physiol. Meas.*, vol. 41, no. 11, 2020, Art. no. 115001.
[12] M. Nikbakht et al., "A synthetic seismocardiogram and electrocardiogram generator phantom," in *Proc. IEEE Sensors*, 2022, pp. 1–4.
[13] M. Chan, L. Klein, J. Fan, and O. Inan, "SCG-RHC: Wearable seismocardiogram signal and right heart catheter database (version 1.0.0), *PhysioNet*, 2023. [Online]. Available: https://doi.org/10.13026/133d-pk11
[14] C. Yang et al., "An open-access database for the evaluation of cardio-mechanical signals from patients with valvular heart diseases," *Front. Physiol.*, vol. 12, 2021, Art. no. 750221.
[15] B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva, "MIMIC-III waveform database (version 1.0)," *PhysioNet*, 2020. [Online]. Available: https://doi.org/10.13026/c2607m
[16] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 48.
[17] J. Skoric, Y. D'Mello, and D. V. Plant, "Synthetic seismocardiography signal generation by a generative adversarial network," in *Proc. IEEE Comput. Cardiol. (CinC)*, Oct. 2023, vol. 50, pp. 1–4.
[18] M. Kaisti, J. Laitala, D. Wong, and A. Airola, "Domain randomization using synthetic electrocardiograms for training neural networks," *Artif. Intell. Med.*, vol. 143, 2023, Art. no. 102583.
[19] J.-P. Sirkiä, T. Panula, and M. Kaisti, "Wearable edge machine learning with synthetic photoplethysmograms," *Expert Syst. with Appl.*, vol. 238, 2024, Art. no. 121523.
[20] "Scipy Gaussian pulse documentation," Accessed: Mar. 12, 2024. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.gausspulse.html
[21] L. Sathyapriya, L. Murali, and T. Manigandan, "Analysis and detection R-peak detection using modified Pan-Tompkins algorithm," in *Proc. IEEE Int. Conf. Adv. Commun. Control Comput. Technol.*, 2014, pp. 483–487.
[22] Y. D'Mello et al., "Real-time cardiac beat detection and heart rate monitoring from combined seismocardiography and gyrocardiography," *Sensors*, vol. 19, no. 16, 2019, Art. no. 3472.
[23] T. Koivisto et al., "Atrial fibrillation detection using MEMS accelerometer based bed-sensor," in *Proc. IEEE Comput. Cardiol. (CinC)*, Piscataway, NJ, USA, Sep. 2019, pp. 1–4.
[24] I. Elnaggar et al., "Multichannel bed based ballistocardiography heart rate estimation using continuous wavelet transforms and autocorrelation," in *2022 Comput. Cardiol. (CinC)*, Piscataway, NJ, USA, 2022, pp. 1–4.