Fetal Heart Rate Estimation in Phonocardiograms Using Deep Learning and Digital Signal Processing

Hugo Carvalho de Moraes
hugo.moraes@icomp.ufam.edu.br
Federal University of Amazonas,
Institute of Computing
Manaus, Amazonas, Brasil

Rafael C. Carvalho
rcc@icomp.ufam.edu.br
Federal University of Amazonas,
Institute of Computing
Manaus, Amazonas, Brasil

Juan G. Colonna
juancolonna@icomp.ufam.edu.br
Victoria University of Wellington
Wellington, New Zealand
Federal University of Amazonas,
Institute of Computing
Manaus, Amazonas, Brasil

Eduardo Freire Nakamura
nakamura@icomp.ufam.edu.br
Federal University of Amazonas,
Institute of Computing
Manaus, Amazonas, Brasil

ABSTRACT

Fetal heart rate monitoring through phonocardiograms enables non-invasive assessment of fetal health, supporting the early detection of potential complications during pregnancy. However, accurate frequency estimation is hindered by various noise sources, such as maternal heart sounds and ambient interference. This study proposes a lightweight 1D-CNN regression model for fetal heart rate estimation, leveraging MFCC and Delta-MFCC coefficients extracted from audio segments. A simple but effective data augmentation technique was also employed to mitigate the scarcity of labeled data. Experiments conducted on the SUFHSDB dataset yielded a mean absolute error of 3.94 ± 0.41 bpm. The results suggest that deep learning-based approaches, especially when combined with data augmentation, are promising alternatives for fetal heart rate estimation, potentially reducing reliance on traditional signal-processing pipelines.

KEYWORDS

Deep Learning Fetal Hearth Rate Digital Signal Processing

1 INTRODUCTION

Fetal heart rate monitoring refers to the monitoring of a fetus' heart rate. This technique allows the assessment of the baby's health and the identification of signs of possible complications. There are monitoring techniques such as phonocardiography (PCG), cardiotocography, electrocardiography (ECG), and magnetocardiography. However, fetal phonocardiography (fPCG) is non-intrusive, low-cost, and suitable for continuous monitoring of fetal heart rate [15].

fPCG is a recording made with equipment that detects mechanical vibrations on the surface of the pregnant woman's abdomen (Figure 1(a)), more specifically in the detection of fetal heart sounds (HS). HS are produced by the closing and opening of the fetal heart valves, and are known as S1 (first heart sound) and S2 (second heart sound). These movements are the main components of the cardiac cycle and can be perceived through fPCG [4].

However, the recordings contain noise that makes it difficult to detect and count HS. Noises include biological effects of the mother's body, such as sounds generated by surrounding organs and tissues, body movements, breathing, uterine contractions, and maternal heart sounds. There are also external noises, such as ambient sounds or interference from other devices. In addition, the fPCG signal varies according to gestational age (stage of pregnancy) or the position of the fetus, which may change during a measurement. The interference caused by these more intense noises overlaps the fetal heartbeat in time and frequency, making it difficult to accurately estimate beats per minute (bpm) [13].

Therefore, if the digital audio signal processing methods and the techniques for estimating bpm were combined and improved, fPCG-based monitoring could become the future of electronic fetal monitoring [7]. This work seeks to explore Deep Learning techniques using convolutional neural networks, residual layers, data augmentation and search for better parameters to treat the data in order to improve and ensure a more accurate and robust estimate of the fetal bpm from its fPCG. Furthermore, this work investigates the impact of the Wavelet and Band Pass filters, widely used in literature, on the proposed model's performance.

This paper is structured as follows: in Section 2, the theoretical framework is presented, addressing the characteristics of fPCG signals, main filtering methods and feature extraction techniques. In Section 3, the related works that use classical algorithms and Deep Learning approaches are detailed. Section 4 describes the adopted methodology, including data preprocessing, segmentation, data augmentation and the proposed model. Then, in Section 5, the experimental results are presented. Finally, Section 6 presents the conclusions and suggestions for future work.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

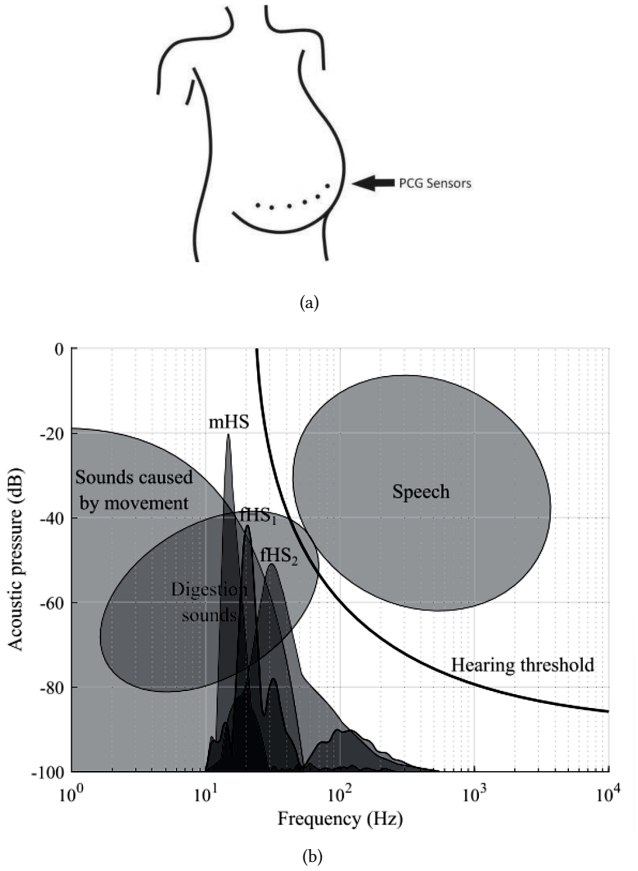

Figure 1: (a) Positioning of PCG sensors [9]. (b) Frequency characteristics of sounds S1 and S2 [13].

2 THEORETICAL FRAMEWORK

2.1 Characteristics of fPCG signals

The acoustic signal of the fetal heart is characterized by low-intensity sounds and a narrow frequency band, resulting from the opening and closing of the heart valves and blood flow, attenuated as it passes through the maternal tissue. In general, S1 has a higher amplitude, lower frequency and longer duration than S2. Fetal heart sounds (fHS) appear in frequency ranges between 20 and 110 Hz, while maternal heart sounds appear between 10 and 40 Hz [13]. The Figure 1(b) illustrates the frequency characteristics of these signals.

Additionally, interfering noises or sounds can be classified as:

- **Sensor and background noise (SNB)**: these random noises are considered 'white' because they cover all frequency bands, significantly altering the signal;
- **Shear noise**: this noise is generated by the movement of the sensor during recording;
- **Ambient noise**: external sounds from the surroundings (conversations, electronic devices, etc.) present in a wide range of frequencies, which can be reduced by adequate positioning of the sensor;
- **Fetal body noises**: these noises can be subdivided into:
  - Fetal respiration (fR): periodic low-frequency signal (0.3 to 1.5 Hz), with moderate intensity;
  - Fetal movements: artifacts from 0 to 25 Hz, including hiccups and respiratory movements, varying in intensity according to fetal activity;
- **Maternal body noises**:
  - Maternal respiration (mR): generally in a low frequency band between 0.2 to 0.5 Hz with relatively constant intensity;
  - Maternal heart sounds: contained in a frequency range varying approximately from 8 to 25 Hz or up to 50 Hz. Generally with greater amplitude than fHS, fR and mR;
  - Uterine contractions: these occur 2 to 5 times every 10 minutes, lasting 15 to 70 seconds, with increasing frequency and intensity as the pregnancy progresses.

2.2 Band pass filter

The Butterworth filter [10] used in this research is an IIR (Infinite Impulse Response) type filter that presents an approximately flat amplitude response in the band pass. This filter is used to attenuate unwanted frequencies while maintaining signal integrity in the fPCG frequency band between 20 and 110 Hz. The transfer function of this filter is expressed as $H(z) = \sum b_k z^{-k} / \sum a_k z^{-k}$, where the coefficients b_k define the impulse response of the filter and a_k influence the stability and behavior of the system.

2.3 Wavelet transform

In this work, Discrete Wavelet Decomposition was used to filter the signal [6]. First, the sign x is represented in the approximation components $A_{2^j}(x) = \sum_k c_{j,k}\phi_{j,k}(x)$ and details $D_{2^j}(x) = \sum_k d_{j,k}\psi_{j,k}(x)$, where 2^j indicates the resolution level, ϕ and ψ are basis functions and $c_{j,k}$ and $d_{j,k}$ are the coefficients associated with the basis functions that in this study were defined by the Coiflet 4 (coif4) wavelet, which has also been used in related works [11–13].

Initially, the signal is decomposed into seven levels using wavelet decomposition. Subsequently, a universal threshold is applied to the detail coefficients, and the signal is then reconstructed with the Inverse Wavelet Transform [12].

2.4 MFCC and Delta MFCC

Mel Frequency Cepstral Coefficients (MFCC) are a compact representation of the spectral characteristics of an audio signal, used in speech recognition and audio processing [5]. The MFCC can be calculated by conducting five consecutive processes, namely signal framing, computing of the power spectrum, applying a Mel filterbank to the obtained power spectra, calculating the logarithm values of all lter banks, and finally applying the DCT.

The idea behind splitting signals into distinct "frames" is to break down the raw data signal into frames where the signal tends to be more stationary. Among the most well-known window functions are the Hanning and Hamming windows [1] which is used in this study. A power spectrum can be described as the distribution of the power of the frequency components that composes the signal. Traditionally, Discrete Fourier Transform (DFT) is utilized to compute the power spectrum, as desbribed below:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi jnk}{N}} \qquad k = 1, 2, 3, \ldots, N-1 \tag{1}$$

where x(n) is discrete signal and N is the length of the signal.

The Mel band-pass filter is a bank of filters, which is constructed based on pitch perception. The Mel filter was originally developed for speech analysis and like human ear perceiving of speech, it targets extracting non-linear representation of the speech signal [1]. The transfer function (TF) of each of the m-th filter can be computed via the equation below:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2}$$

where, $f(m)$ is the center frequency of the triangular filter and $\sum_m H_m(k) = 1$. The Mel scale to the response frequency and vice versa is computed by Equations 3 and 4:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{3}$$

$$f = 700 \left(10^{\frac{m}{2595}} - 1\right) \tag{4}$$

A Discrete Cosine transform (DCT) expresses a finite sequence of data points regarding a summation of cosine functions oscillating at different frequencies In the MFCC process, the DCT is applied on the Mel filter bank to select most accelerative coefficients or to separate the relationship in the log spectral magnitudes from the filter-bank [1]. The DCT is computed by the Equation 5.

$$X(k) = \sum_{n=0}^{N-1} x_n \cdot \cos\left(\frac{2\pi jnk}{N}\right), \quad k = 1, 2, 3, \ldots, N-1 \tag{5}$$

where x_n is a discrete signal and N is the length of the signal.

Intuitively, the first coefficients carry more global information of the spectrum, while the last ones capture finer details. The first coefficient C_0 represents the total energy of the signal. This coefficient is often removed in some applications because it represents a general characteristic of the signal loudness rather than capturing timbre information [16], so it's worth testing this approach in this study. Delta MFCCs arise from taking the first derivative of the MFCCs with respect to time, which allows us to represent the temporal rate of change of the MFCCs, capturing the dynamics of the signal. The derivative can be simply obtained as $\Delta\, C^n = C^n_{t+m} - C^n_t$, where n is the nth coefficient and $t + m$ is the coefficient in the next time window.

3 RELATED WORKS

3.1 Approaches with Classical Algorithms

The first approaches to analyze fPCG audio in the literature used classical audio processing techniques. These works were important to identify signal characteristics and determine more appropriate noise filters. In order for detection to work with a classical approach, the fPCG signals must be filtered and noise-free; otherwise, more detection errors will occur.

For example, the authors of PCG-Delineator [10] proposed to filter the fPCG signal using the Wavelet transform with 4th order Coiflets, with 7 levels of decomposition and smooth thresholding, before proceeding with the peak counting algorithm. In the bpm estimation step, an iterative threshold-based algorithm is employed to detect the peaks of the S1 and S2 sounds. Initially, the S1 detection procedure establishes an amplitude threshold corresponding to 30% of the maximum amplitude of the filtered fPCG signal, with the condition that there is at least 40 ms of separation between two consecutive S1 peaks. This temporal characteristic allows discarding possible detections without physiological significance (false positives).

The first step, identify all potential S1 peaks; then, the distances between them are calculated to perform an additional temporal control, aiming to identify the peaks that were not detected in the initial evaluation. After all S1 peaks have been located and confirmed, the search for S2 peaks is performed [10]. In non-pathological cases, S2 occurs shortly after S1 in a normal cardiac cycle. Consequently, the S2 identification procedure, also based on an amplitude threshold, is performed after S1 detection. Specifically, S2 identification is based on the premise that the diastolic duration (i.e., the time interval between S2 and S1 sounds) is longer than the systolic duration (i.e., the interval between S1 and S2 sounds). Thus, S2 must occur at least 100 ms after the previous S1 and at most 200 ms before the next S1. Furthermore, S2 must have an amplitude less than 80% of that of the previous S1.

The AdvFPCG-Delineator [12] uses the same noise filter based on the wavelet transform with 4th order Coiflets. However, unlike the PCG-Delineator, the AdvFPCG applies two previous steps: (1) the normalization of the maximum and minimum amplitudes of the signals to vary between ±100; and (2) the application of a 6th order bidirectional Butterworth filter, band pass type, with lower and upper cutoff frequencies of 20 Hz and 120 Hz. The successive application of these two combined filtering techniques contributes to obtaining a cleaner fPCG signal scalogram. Finally, the identification of S1 and S2 is performed in the scalogram, using physiological properties of fPCG and the same iterative algorithm employed in PCG-Delineator.

These two classical signal processing approaches establish a noise filter technique that can be incorporated as a pre-processing step in more modern algorithms based on Deep Learning. This work proposes such an approach in Section 4.

Although existing methods demonstrate promising results, they exhibit several limitations. One notable issue is the inability to detect cardiac arrhythmias that produce abrupt pulses, as the algorithm enforces a minimum interval of 100 ms between consecutive S1 sounds. Additionally, while false-positive peak detections may occur, this effect is mitigated by averaging S1 intervals; however, this approach requires a cleaner phonocardiogram signal to minimize excessive false positives. Another challenge lies in the initial detection of the first S1 sound. Since the audio signal may contain residual oscillations that do not fully decay to zero, relying solely on a time threshold can lead to erroneous peak identification.

Figure 2 illustrates the time-threshold algorithm [12] through two graphical representations. The first graph depicts a heartbeat

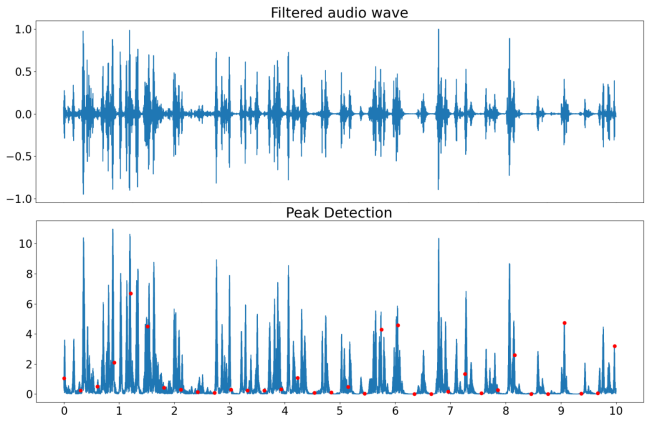

Figure 2: Reproduction of time-threshold algorithm using a graph, as described in [12], using a fPCG recording captured under realistic noise conditions. The maximum peaks detections demonstrate the failures of this temporal approach under non-ideal circumstances.

audio signal, with amplitude normalized to the interval $[-1, 1]$, plotted over time. The second graph presents the corresponding signal after undergoing a wavelet transform, represented as the sum of amplitude values across all frequencies over time. Superimposed on this transformed signal are red markers, indicating the peaks identified by the time-threshold algorithm. It is evident that the peak detection is incorrect: in this example, the fetal heart rate should have been 138 bpm, but the algorithm returned 198.58 bpm. Therefore, our aim is to develop a more robust, noise-tolerant algorithm—such as one based on deep learning.

3.2 Approaches with Deep Learning

In modern literature, two papers have addressed the problem of fetal heart rate estimation using Deep Learning, FHSU-Net [2] and FHSU-NetR [3]. Both works used the U-Net architecture to filter noise and extract the clean fetal signal, aiming to approximate ECG data.

FHSU-Net is a deep learning model based on the U-Net [8] architecture, specifically designed to extract fetal heart sounds directly from abdominal phonocardiograms without preprocessing steps, noise filtering, or manual signal adjustments. The framework incorporates encoder and decoder modules with one-dimensional (1D) convolutional layers, taking advantage of the robustness of U-Net to input data perturbations. The training data consisted of simultaneous PCG and ECG recordings of 20 healthy pregnant women, captured by piezoelectric sensors and noninvasive electrodes, with 16-bit resolution and a 1 kHz sampling rate. The signals were temporally aligned during preprocessing, and noisy or uninformative channels were discarded based on a cross-correlation analysis (threshold of 0.3). Masks based on the R-peaks of the fetal ECG served as reference for training. The model achieved an average error of 5.18 bpm in estimating fetal heart rate.

Despite the promising results, the authors of FHSU-Net identified that the small size of the dataset limited the model's performance. To overcome this limitation, they proposed FHSU-NetR (Fetal Heart Sounds U-Net Transformer), which combines three U-Net architectures to separate maternal respiratory signals, fetal and maternal heart sounds, in addition to incorporating Transformer layers adapted for time series. FHSU-NetR was pre-trained with 380,000 synthetic samples, generated from signal separation techniques and beat shapes extracted from fECG. In the evaluation, the model presented an average error of 1.5 ± 18.13 bpm when comparing fPCG with fECG (ground truth), highlighting a critical limitation, the exceptionally high standard deviation (18.13 bpm), a severe inconsistency in prediction accuracy likely due to dataset limitations and synthetic data biases. However, the hybrid U-Net-Transformer architecture is complex with many convolution layers and transformer connections that connect to other outputs and therefore depends on a lot of data to be trained, which makes its clinical validation difficult due to the lack of sufficiently broad real bases.

These approaches require specialized equipment to acquire synchronized fPCG and ECG signals, as well as a human specialist to label the data, making them very expensive. Our work eliminates the need for synchronized ECG data and thus cannot be directly compared to FHSU-Net.

4 METHODOLOGY

This work proposes a model to extract fetal heart rate directly from phonocardiogram signals. The approach uses one-dimensional convolutional neural networks (1D-CNN) with residual connections for regression. Training, testing, and validation data were obtained from the Shiraz University Fetal Heart Sounds Database (SUFHSDB), which provides bpm labels for 10-second audio segments. All experiments were performed with the Python programming language, using the libraries LIBROSA[1] and KERAS[2]. A version of the scripts used in the experiments are available in the repository https://github.com/AnonPaperSub27/FPCG-Residual-1D-CNN.

4.1 Dataset

The availability of bpm-labeled fPCG recordings is critical for this research. Public datasets in the *PhysioNet*[3] are limited, mostly unlabeled, and some include simulated signals. Therefore, the SUFHSDB database [9] was chosen, as it is the only database with explicit labels of fetal bpm and clinical metadata of patients. This database contains fPCG recordings of 109 pregnant women (including 7 twin cases), collected at Hafez Hospital (Iran) with a digital stethoscope positioned on the lower abdominal region, with participants aged between 16 and 47 years. There are 119 recordings, 99 with a single signal, 3 with two signals (recorded twice) and 7 pairs of twins. The signals were captured in 16 bits, with a sampling rate of 16 kHz. Each recording was segmented into 10-second windows labeled with the corresponding fetal bpm, resulting in 621 temporal windows. The Figure 3(a) shows the distribution of labels among the examples in this database. It is observed that fetuses have an average bpm close to 140.

4.2 Preprocessing

Preprocessing involved the following steps:

[1]https://librosa.org/
[2]https://keras.io/
[3]https://physionet.org/

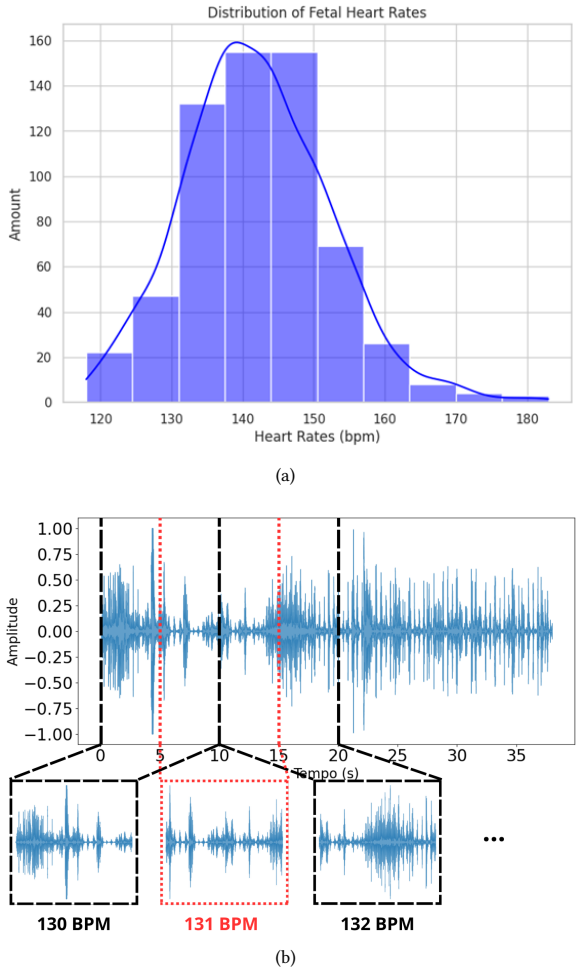

Figure 3: (a)Histogram showing the distribution of bpm across the labels of the examples in the SUFHSDB database. (b) Audio data augmentation technique, in red is the new data and in black the original labeled segments.

(1) **Normalization**: signal amplitude adjusted to the range $[-1, 1]$;
(2) **Band pass filter**: application of a 6th order Butterworth filter (20–120 Hz) to isolate the typical spectral range of fetal signals [13];
(3) **Wavelet decomposition**: The 4th order Coiflets wavelet with 7 decomposition levels was used, followed by the calculation of the universal threshold to reduce residual noise and reconstruct the signal from the processed detail coefficients;
(4) **Segmentation**: division of the preprocessed signal into 10-second windows, aligned to the bpm labels for subsequent data augmentation.

The wavelet filtering and decomposition steps were inspired by established signal processing techniques [10, 12].

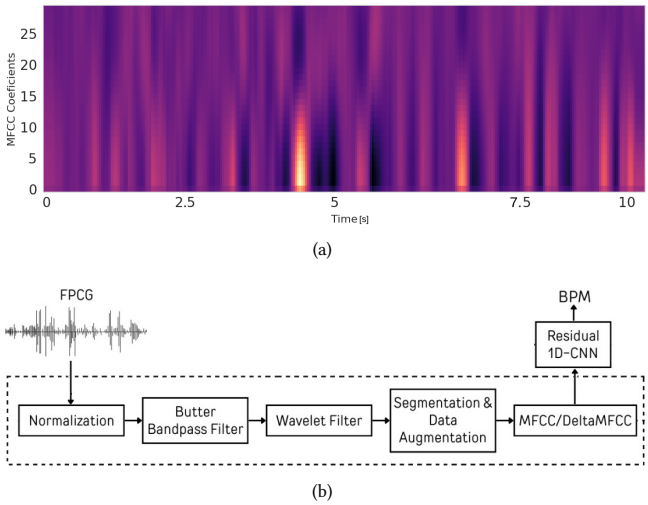

Figure 4: (a) MFCC spectrogram of a 10 s segment of the fPCG dataset. The color intensity represents the values of the 30 MFCC coefficients at each time point. (b) Illustration of the preprocessing pipeline.

4.3 Data Segmentation and Augmentation

Due to the small amount of labeled data available, a data augmentation method was used. Since each labeled instance is divided into non-overlapping 10-second segments, the last 5 seconds of one segment and the first 5 seconds of the next segment were removed to create a new segment whose label is the average of the labels of the original segments, as illustrated in Figure 3(b). After data augmentation, 513 new instances were generated from the 621 original time windows, totaling 1,134 examples, which corresponds to an increase of 82.6%. Finally, segments with a duration of less than 10 seconds are discarded, since, in the database, the audios do not have a fixed duration, resulting in fractions of windows that do not have a label.

No augmented data was used for tests, since this data could artificially change the metrics.

4.4 MFCC Spectrogram

After the preprocessing, segmentation and data augmentation steps, 30 MFCCs were extracted using a 1024-point FFT and a hop length of 512 points. Thus, considering a sampling frequency of 16 kHz, for each 10s audio segment, there are 313 FFTs organized in a matrix of dimension $\mathbb{R}^{30\times313}$. This matrix, called MFCC spectrogram, is illustrated in Figure 4(a). By calculating the difference between each successive column of this matrix, obtaining the first-order Delta-MFCCs; by repeating the calculation of the difference, obtaining the second-order Delta-MFCCs, and so on. Delta-MFCCs are useful to characterize the instantaneous variation of the signal, better highlighting abrupt temporal changes. Finally, the generated matrix is transposed, since the input layer of the 1D-CNN model requires the temporal sequence to be arranged in the rows of the matrix.

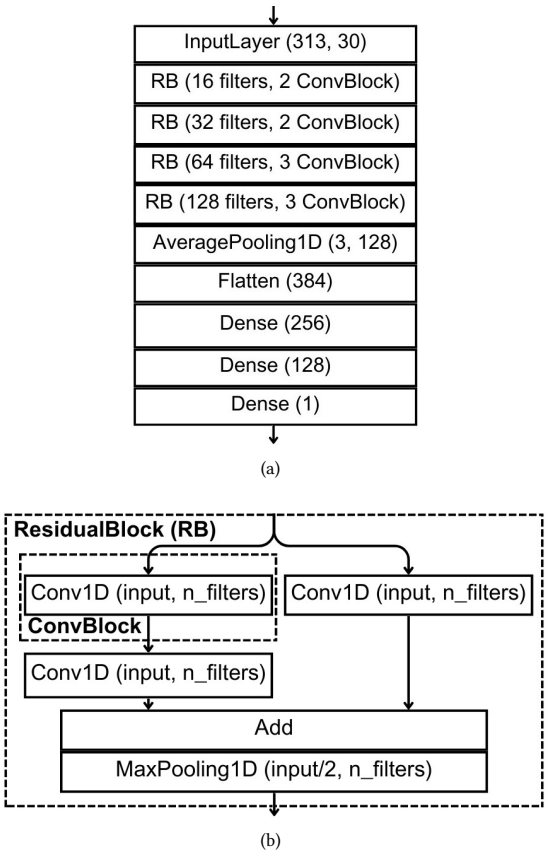

Figure 5: (a) Architecture of the developed model and (b) The residual block (RB) used. The model includes four RBs and a matrix input of 313 temporally ordered vectors with 30 MFCCs each, forming an MFCC spectrogram of 313×30**.**

For the experiments, the following variables were evaluated: (i) the order of the Delta-MFCC, that can be zero (static MFCC), one (show velocity of the MFCC coefficients), or two (show how much the MFCC coefficients accelerate); (ii) the alternation between the normalization modes, i.e., normalizing all coefficients together(all), by getting the max value of the matrix and dividing all values by it or normalizing the zero coefficient separately (separated), using the max value of the zero coefficient for the other values on the same zero coefficient and get tha max value from the rest and dividing the same rest by it; and (iii) the option of removing or not the zero coefficient (energy coefficient) as explained in the section 2.4. Finally, the preprocessing pipeline is illustrated in Figure 4(b).

4.5 Model

The proposed model is based on the *speaker recognition* from the Keras library, designed for human speech recognition and is robust to background noise and variations in recording quality. After adaptations, the model has 468,497 trainable parameters (1.79 MB) and uses an input with dimension 313×30. The activation functions are ReLU, except in the last layer, which uses sigmoid activation for regression. The Adam optimizer was configured with a learning rate of 0.001, and the loss function chosen was the mean squared error (MSE). Figure 5 illustrates the proposed model along with the values of the adopted hyperparameters.

The main backbone of this architecture is composed of 4 residual blocks, followed by an *AveragePooling1D* layer, a *Flatten* layer and three *Dense* layers. The composition of the residual blocks with skip connection is illustrated in Figure 5(b). These residual blocks help the network to identify important features of the input, which should be passed on to the output layer.

4.6 Training and validation

For training, the *EarlyStopping* callback was used, which stops execution if the validation metric does not show improvement after ten epochs. The data were split into 80% for training and validation and 20% for testing. In order to avoid information leakage and artificial elevation of the metric results, time windows of the same patient in the training and validation sets were not mixed. Finally, a cross-validation was performed, with 10 sets (*10-CV*) to evaluate each experimental configuration presented in Table 1.

The quality of the trained model was assessed using the mean absolute error (MAE) in the five cross-validation sets. In addition to the MAE, the respective standard deviations were also calculated. The MAE is a metric that assesses the accuracy of a regression model by calculating the average of the absolute errors between the real values (y_i) and the predicted values ($\hat{y}i$), according to MAE = $1/n \sum_{i=1}^{n} |y_i - \hat{y}_i|$, where n is the total number of observations. A lower MAE value indicates a lower error in the estimation of fetal bpm.

The main advantage of using MAE is its unit, this metric shows exactly how many bpm the prediction is off without any extra calculations, making the results easier to understand and check.

5 RESULTS

The results of the experiments for the various experimental configurations of variables, with and without data augmentation, are found in Table 1. To test the impact of the band pass and wavelet filters, discussed respectively in Section 2.2 and 2.3, the table also includes results for the same experimental configurations, without these filters

Analysis of the mean absolute errors (MAE) indicates that the best performance with filters was achieved when using Delta equal to 1, separate normalization of the coefficients, and inclusion of the energy coefficient (C_0). Under this configuration, the MAE was 4.85 ± 1.03 bpm without data augmentation and 4.15 ± 0.41 bpm with augmentation. In the absence of filters, the MAE was 4.93 ± 1.54 without augmentation and 4.07 ± 0.46 with augmentation, showing a slight improvement with the augmentation technique.

Notably, removing the filters and using Delta equal to 1 without the energy coefficient (C_0), in combination with data augmentation, resulted in the lowest MAE: 3.94 ± 0.41 bpm. Although the mean MAE remains within the variance of the same filtered configuration, the variance is less than half, indicating significantly greater stability in predictions.

Table 1: Results of the trained model for the different experimental settings. Split normalization can only be applied when C_0 is present.

Filters	Delta	Normalization	Have C_0?	MAE±var (No Augmentation)	MAE±var (With Augmentation)
With	0	separated	yes	5.34 ± 1.70	4.61 ± 1.10
	0	all	yes	7.68 ± 0.38	7.64 ± 0.33
	1	**separated**	**yes**	**4.85 ± 1.03**	**4.15 ± 0.41**
	1	all	yes	4.75 ± 0.97	4.18 ± 0.32
	2	separated	yes	4.62 ± 0.39	5.13 ± 1.40
	2	all	yes	5.26 ± 1.46	5.09 ± 1.44
	0	-	no	5.42 ± 1.61	4.54 ± 1.19
	1	-	no	5.04 ± 1.33	4.49 ± 1.00
	2	-	no	4.73 ± 0.41	4.69 ± 0.44
Without	0	separated	yes	5.97 ± 1.72	5.47 ± 1.78
	0	all	yes	7.68 ± 0.33	6.90 ± 1.75
	1	separated	yes	4.93 ± 1.54	4.07 ± 0.46
	1	all	yes	4.56 ± 1.07	4.59 ± 1.67
	2	separated	yes	5.64 ± 1.48	4.55 ± 1.32
	2	all	yes	6.57 ± 1.82	5.87 ± 2.02
	0	-	no	5.97 ± 1.72	5.47 ± 1.78
	1	**-**	**no**	**4.90 ± 1.38**	**3.94 ± 0.41**
	2	-	no	6.46 ± 1.80	5.19 ± 1.57

Table 2: Wilcoxon Signed-Rank test results comparing the configurations below with and without filters.

Configuration	p-value
Delta=1, Normalization=separated, HaveC_0=yes, With Augmentation	0.0032
Delta=1, Normalization=None, HaveC_0=no, With Augmentation	4.5×10^{-10}

To rigorously evaluate whether the removal of filters leads to improved model performance, statistical analyses were performed using the Wilcoxon Signed-Rank Test [14] on the two best-performing configurations. The resulting p-values are presented in Table 2. Under the null hypothesis that the paired samples (i.e., results with and without filters for each configuration) are drawn from the same distribution, the p-values for both configurations were found to be below 0.01. This enables rejection of the null hypothesis at the 99% confidence level, indicating that the inclusion of filters significantly degrades the model's performance.

The observed degradation in model performance following the application of filters suggests that relevant components of the audio signal may have been inadvertently removed. These components could be unrelated to the target variable (BPM), potentially representing characteristic background noise present in the dataset that unintentionally aids model learning. Alternatively, they may correspond to physiologically meaningful elements of the fetal heart sound itself, indicating that the filtering process is attenuating information essential for accurate prediction.

The convergence of the loss function curves (MSE) during the training of the model with the best result can be observed in Figure 6, which presents the mean MSE of the *10-folds* together with the variance, indicated by the shaded area.

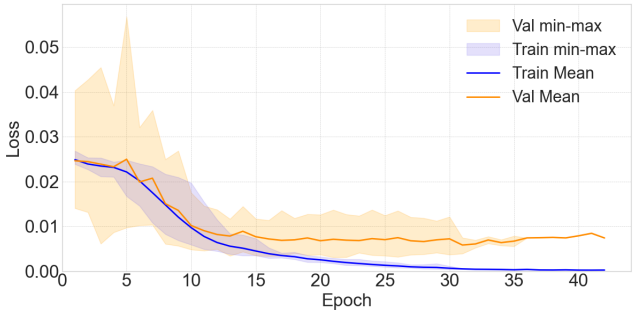

Figure 6: Average of the training curves of the folds of the best combination of variables.

In general, the data augmentation technique proved to be advantageous in almost all scenarios evaluated, which is justified by the greater number of examples used in training the model. Although the method of generating new instances is simple, the results indicate that it offers concrete benefits for the bpm regression task. The comparison between the approaches with and without the energy coefficient is inconclusive, as there is no consistency in

evidencing improvements in the results with its inclusion. It is also observed that, when Delta is equal to zero and the energy coefficient is removed, normalization generates a significant difference in the results.

6 CONCLUSION

This work investigates the use of a lightweight 1D-CNN-based regression model to estimate fetal heart rate (FHR, in bpm) from fetal phonocardiogram (fPCG) signals. Unlike traditional approaches that rely on explicit detection of S1 and S2 components (first and second heart sounds), we propose a straightforward solution for bpm prediction, integrating the following steps:

(1) Band-pass filtering for spectral isolation;
(2) Wavelet decomposition for noise removal in the band of interest (20–120 Hz);
(3) Extraction of Delta-MFCC coefficients to capture the signal's temporal dynamics;
(4) A simple but effective technique for data augmentation;
(5) Training of a 1D temporal convolutional neural network adapted from a speech recognition application.

This approach builds upon the methods described in Section 3, which utilize band-pass and wavelet filters commonly adopted in acoustic signal analysis. To assess their impact, the model was also trained without these filters. As shown in Table 1, performance improved in the absence of filtering. The Wilcoxon Signed-Rank test confirmed that this improvement is statistically significant, rejecting the null hypothesis that filters have no effect. Therefore, we conclude that filtering degrades the performance of the proposed model, and that a deep learning-based approach can outperform classical acoustic analysis frameworks, even without pre-processing.

Overall, this study demonstrates that a deep learning model can serve as an effective alternative to traditional phonocardiogram analysis pipelines. Notably, the proposed model achieves higher performance, greater robustness, and lower computational cost when used without pre-processing filters.

The results indicate satisfactory performance, particularly highlighting the impact of the data augmentation strategy, which significantly reduced the mean absolute error (MAE). The best model (without filters) achieved an MAE of 3.94 bpm. Considering the average FHR is approximately 140 bpm (as shown in Figure 3(a)), this corresponds to an average error of roughly ±4 bpm, or 2.8% of the mean FHR.

For comparison, the FHSU-Net method reported an average error of 5.18 bpm, whereas our proposed model achieved a lower MAE of 3.94 bpm (±0.41). More importantly, our model demonstrated significantly greater stability, with a much smaller standard deviation compared to the FHSU-Net's exceptionally high variability of ±18.13 bpm.

Direct comparison with related works is challenging due to limitations in methodological transparency, including the unavailability of tools, filter parameters, dataset labels, and use of non-public databases. Furthermore, the reported results often employ different evaluation metrics, further hindering reproducibility and comparison. In this context, our work has the potential to serve as a reliable baseline for future regression-based approaches to fetal heart rate estimation.

To improve the proposed method, future work should consider: (i) exploring hybrid architectures (e.g., CNN-Transformer) to capture long-term dependencies and incorporate longer temporal windows; (ii) developing data augmentation strategies that simulate realistic conditions (e.g., respiratory noise, fetal movement); and (iii) validating model generalization on Brazilian hospital datasets, including pathological cases such as fetal tachycardia and intrauterine growth restriction.

ACKNOWLEDGMENTS

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 39 of Decree n°10.521/2020, was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n°8.387/1991 through agreement 001/2020, signed with UFAM and FAEPI, Brazil. This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (AUXPE-CAPES-PROEX), Financing Code 001, and by the Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM, through the PDPG/CAPES project.

REFERENCES

[1] Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. 2022. Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access* 10 (2022), 122136–122158. https://doi.org/10.1109/ACCESS.2022.3223444
[2] Mohanad Alkhodari, Murad Almadani, Samit Kumar Ghosh, and Ahsan H. Khandoker. 2023. Fhsu-Net: Deep Learning-Based Model for the Extraction of Fetal Heart Sounds in Abdominal Phonocardiography. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. 1–6. https://doi.org/10.1109/MLSP55844.2023.10285907
[3] Murad Almadani, Mohanad Alkhodari, Samit Ghosh, and Ahsan Khandoker. 2023. FHSU-NETR: Transformer-based Deep Learning Model for the Detection of Fetal Heart Sounds in Phonocardiography. https://doi.org/10.22489/CinC.2023.026
[4] Vijay S. Chourasia, Anil Kumar Tiwari, and Ranjan Gangopadhyay. 2014. A novel approach for phonocardiographic signals processing to make possible fetal heart rate evaluations. *Digital Signal Processing* 30 (2014), 165–183. https://doi.org/10.1016/j.dsp.2014.03.009
[5] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. 2004. Speaker identification using mel frequency cepstral coefficients. *variations* 1, 4 (2004), 565–568.
[6] S.G. Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 7 (1989), 674–693. https://doi.org/10.1109/34.192463
[7] Radek Martinek, Jan Nedoma, Marcel Fajkus, Radana Kahankova, Jaromir Konecny, Petr Janku, Stanislav Kepak, Petr Bilik, and Homer Nazeran. 2017. A Phonocardiographic-Based Fiber-Optic Sensor and Adaptive Filtering System for Noninvasive Continuous Fetal Heart Rate Monitoring. *Sensors (Basel)* 17, 4 (Apr 2017), 890. https://doi.org/10.3390/s17040890 arXiv:http://www.mdpi.com/1424-8220/17/4/890/pdf The authors declare no conflict of interest..
[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *LNCS* 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
[9] Maryam Samieinasab and Reza Sameni. 2015. Fetal phonocardiogram extraction using single channel blind source separation. *2015 23rd Iranian Conference on Electrical Engineering* (2015), 78–83. https://api.semanticscholar.org/CorpusID:23679962
[10] Annachiara Strazza, Agnese Sbrollini, Valeria Battista, Rita Ricci, Letizia Trillini, Ilaria Marcantoni, Micaela Morettini, Sandro Fioretti, and Laura Burattini. 2019. PCG-Delineator: an Efficient Algorithm for Automatic Heart Sounds Detection in Fetal Phonocardiography. https://doi.org/10.22489/CinC.2018.045
[11] Annachiara Strazza, Agnese Sbrollini, Marica Olivastrelli, Agnese Piersanti, Selene Tomassini, Ilaria Marcantoni, Micaela Morettini, Sandro Fioretti, and Laura Burattini. 2019. *PCG-Decompositor: A New Method for Fetal Phonocardiogram Filtering Based on Wavelet Transform Multi-level Decomposition.* 47–53. https://doi.org/10.1007/978-3-030-31635-8_6
[12] Selene Tomassini, Agnese Sbrollini, Annachiara Strazza, Reza Sameni, Ilaria Marcantoni, Micaela Morettini, and Laura Burattini. 2020. AdvFPCG-Delineator: Advanced delineator for fetal phonocardiography. *Biomedical Signal Processing and Control* 61 (2020), 102021. https://doi.org/10.1016/j.bspc.2020.102021
[13] Radana Vilimkova Kahankova, Martina Mikolasova, Rene Jaros, Kateřina Barnová, Martina Ládrová, and Radek Martinek. 2022. A Review of Recent Advances and

Future Developments in Fetal Phonocardiography. *IEEE Reviews in Biomedical Engineering* PP (06 2022), 1–1. https://doi.org/10.1109/RBME.2022.3179633

[14] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.

[15] Pengjie Zhang, Shiwei Ye, Zhipei Huang, Dina Jiaerken, Shuxia Zhao, Lingyan Zhang, and Jiankang Wu. 2019. A Noninvasive Continuous Fetal Heart Rate Monitoring System for Mobile Healthcare Based on Fetal Phonocardiography. In *Advances in Body Area Networks I*, Giancarlo Fortino and Zhelong Wang (Eds.). Springer International Publishing, Cham, 191–204.

[16] Bin Zhen, Xihong Wu, Zhimin Liu, and Huisheng Chi. 2001. On the Importance of Components of the MFCC in Speech and Speaker Recognition. *Acta Scientiarum Naturalium-Universitatis Pekinensis* 37, 3 (2001), 371–378.