Complexity-Reduced End-to-End Fetal ECG Signal Recovery and QRS Complex Detection

Julia C. Remus[1], **Thiago L. T. da Silveira**[1,2]

[1]Institute of Informatics – Universidade Federal do Rio Grande do Sul
Porto Alegre – RS – Brazil
Department of Applied Computing – Universidade Federal de Santa Maria
Santa Maria – RS – Brazil

juliacremus@gmail.com, thiago@inf.ufsm.br

Abstract. *Non-invasive electrocardiogram (niECG) enables pregnancy monitoring by assessing the mother's and fetus' health through maternal abdominal signal acquisition. However, isolating fetal ECG (fECG) is challenging due to low signal-to-noise ratio and time and frequency overlap with maternal cardiac activity. Prior studies focused on fetal R-peak detection and more recent works target full fECG waveform noise-free reconstruction, but at a high computational cost, precluding real-world applications. This paper explores different ways of reducing the complexity of state-of-the-art deep learning-based method for fECG recovery and fetal QRS complex localization. Results indicate that one encoder-decoder block pair can be removed without significantly impacting the metrics. While the other complexity-reduction options yield slightly lower metrics compared to the baseline and related works, they can be fine-tuned and adapted based on the specific requirements and objectives of the application.*

1. Introduction

Access to healthcare reduces maternal mortality, but disparities persist worldwide based on region and income, hindering progress toward Sustainable Development Goal 3.1 [UNICEF 2019, WHO]. During pregnancy, a key concern is the fetus' oxygen deficiency, which can cause severe complications or death. Fetal oxygenation can be monitored through the analysis of heart rate (fHR) and mother's uterine contractions (UC) [Vullings et al. 2009]. Assessing the fetus' oxygenation can prevent stillbirths or short and long-term injuries by enabling fast medical decisions [Mendis et al. 2023]. Cardiotocography (CTG) is a widely used technique for simultaneously monitoring fHR and UC. However, CTG has not contributed to reducing fetal mortality, and it is correlated with the increase in C-section rates due to misinterpretation [Kahankova et al. 2020]. As a promising avenue, non-invasive ECG (niECG) provides detailed cardiac data more easily, enabling remote monitoring, particularly for high-risk pregnancies [Sameni 2021].

NiECG enables pregnancy assessment through maternal abdominal ECG (aECG) acquisition and analysis. AECG signals are composite signals containing maternal ECG (mECG), fetal ECG (fECG), UC artifacts, and environmental noise [Kahankova et al. 2020]. Though niECG allows cardiac health analysis for both mother and fetus, reliable fECG signal extraction remains challenging due to (i) temporal and spectral overlap between maternal and fetal QRS complexes and (ii) low fECG signal-to-noise ratio, particularly during early gestation ($<$ 20 weeks) [Kahankova et al. 2020].

The QRS complex – a “curve” that represents ventricular depolarization – contains critical features such as the R-peak markers used for heart rate calculation in adult and fetus signals [Jezewski et al. 2012a]. A way to obtain accurate fECG signals is measuring cardiac activity by placing electrodes on his/her scalp (sECG). However, this signal acquisition strategy can only be applied during labor, leads to very noisy signals, and is prone to medical contamination [Behar et al. 2016].

To recover the fECG signal from aECG signal, different methods were proposed in the literature [Behar et al. 2016, Kahankova et al. 2020]. Among them, deep learning-based methods have shown high accuracy in fetal QRS (fQRS) complex detection and started paying attention to the retrieval of morphologically accurate fECG signal from the aECG signal and mimicking a pre-processed sECG signal [Rahman et al. 2023, Zhong et al. 2019, Ghonchi and Abolghasemi 2022, Mohebbian et al. 2022, Wang et al. 2023, Barnova et al. 2024]. Recently, [Remus and da Silveira 2024] proposed an end-to-end solution tackling both fECG recovery and fQRS complex detection, showing improved results in the detection compared with the gold standard algorithm. Nevertheless, the authors' design choices lead to a model with increased size and complexity. Searching for neural networks with high-accuracy and reduced computation time is a prominent effort towards efficient signal processing applications for low-power devices [Liu et al. 2025].

Deep learning models like the ones from [Remus and da Silveira 2024, Zhong et al. 2019, Rahman et al. 2023, Ghonchi and Abolghasemi 2022] comprise different building blocks (encoder and decoder blocks) that might contribute more or less to the target task(s). We assess these blocks' importance for the recent approach from [Remus and da Silveira 2024], later called baseline, and evaluate the use of parameter-free approaches for signal reconstruction. Our methodology emphasizes computational efficiency while keeping high quality in fECG signal recovery and fetal QRS complex detection, targeting potential real-time in-device implementation. In this work, we perform a throughout evaluation considering inter-subject and cross-dataset analysis in two datasets [Matonia et al. 2020, Sober and Marco 2007], comparing results with state-of-the-art works [Remus and da Silveira 2024, Mohebbian et al. 2022, Ghonchi and Abolghasemi 2022, Rahman et al. 2023, Wang et al. 2023].

The rest of this paper is organized as follows. Section 2 discusses related works. Section 3 describes the proposed methodology. The results are presented and discussed in Section 4. Section 5 draws the conclusions of this work and points out future direction.

2. Related Work

Deep learning-based methods have shown to produce high accuracy in detecting fQRS complexes and the potential to retrieve morphologically accurate fECG signals from aECG for interpretation [Rahman et al. 2023, Mohebbian et al. 2022, Zhong et al. 2019, Ghonchi and Abolghasemi 2022, Shi et al. 2023, Remus and da Silveira 2024]. Available models for fECG signal retrieval are founded on convolutional neural networks (CNN) and architectures like ecnoder-decoders (ED) [Rahman et al. 2023, Zhong et al. 2019, Ghonchi and Abolghasemi 2022, Remus and da Silveira 2024] or generative adversarial networks (GANs) [Mohebbian et al. 2022, Wang et al. 2023].

The work from Zhong et al. [Zhong et al. 2019] introduced an ED coupled with a fully-connected layer at the end and used $\tanh(\cdot)$ as an activation function. Their model was trained in a synthetic dataset [Andreotti et al. 2016] and evaluated in the ADFECG [Jezewski et al. 2012b] and PhysioNet 2013 [Silva et al. 2013] datasets. Ghonchi and Abolghasemi [Ghonchi and Abolghasemi 2022] proposed a model that uses two different attention-based masks coupled with a bi-directional Long Short-Term Memory network in the latent space. Finally, Rahman et al. [Rahman et al. 2023] presented a modified LinkNet structure with deep supervision and dense blocks to improve generalization. They trained and evaluated their method in two datasets [Jezewski et al. 2012b, Matonia et al. 2020], highlighting the morphologically accurate fECG signal they retrieve. Mohebian et al. [Mohebbian et al. 2022] proposed a GAN coupled with an attention head, showing promising results with a few-parameter model. More recently, Wang et al. [Wang et al. 2023] proposed a CycleGAN-like model, changing the generator to an auto-encoder with cross-correlation and residual connections between the encoder and decoder branches. Generative models offer a potential solution for the fECG signal recovery task, but they might suffer from hallucination, where the model generates plausible but incorrect content, or the inability to generalize across datasets [Cohen et al. 2024]. All the works revised above pre-process the aECG and sECG signals and use external tools [Pan and Tompkins 1985] to detect the fQRS from the generated fECG signal.

Unlike previous works, Remus and da Silveira [Remus and da Silveira 2024] proposed an end-to-end, noise-aware CNN-based ED for both fECG signal recovering and fQRS complex detection. Their method relies on the defining regions of interest (RoIs) based on the fetal R-peak annotations and a custom loss function that balances fECG signal recovering, noise filtering, and fQRS complex localization. Their model is trained in a supervised manner, receives multichannel aECG signals as input and use the sECG signal and fQRS complexes modeled as Gaussians as ground-truth. The architecture from Remus and da Silveira [Remus and da Silveira 2024] comprises a single shared encoder and two specialized decoder branches. The decoder branches return the recovered fECG signal and the RoI-based fQRS complex localization.

The method from Remus and da Silveira [Remus and da Silveira 2024] is innovative for tackling the fECG signal recovery and fQRS complex detection at once and does not require pre- or pos-processing. As they provide open-source implementation[1], one can assess the impact of each module employed to the end tasks quality, model size, and inference time – opening room for further investigation. Their model has more than 20 million parameters, representing 80 Mb [Remus and da Silveira 2024]. Focusing on low-cost implementations, we evaluate options to reduce the model size by combining the two branches from the baseline model, removing encoder and decoder blocks and replacing the transpose convolution by one parameter-free interpolation method to decode the signal. Previous studies have emphasized the need for reduced computational resource usage, particularly in edge inference, where achieving low latency and enhanced cybersecurity is possible, both critical factors in healthcare applications [Shuvo et al. 2023].

3. Proposed Methodology

To optimize the baseline RoI-based model [Remus and da Silveira 2024] – a five-block ED with a single encoder and two decoder branches –, we evaluate computational cost re-

[1] https://github.com/tlts-lab-ufrgs/fecg-research-roi-ED.

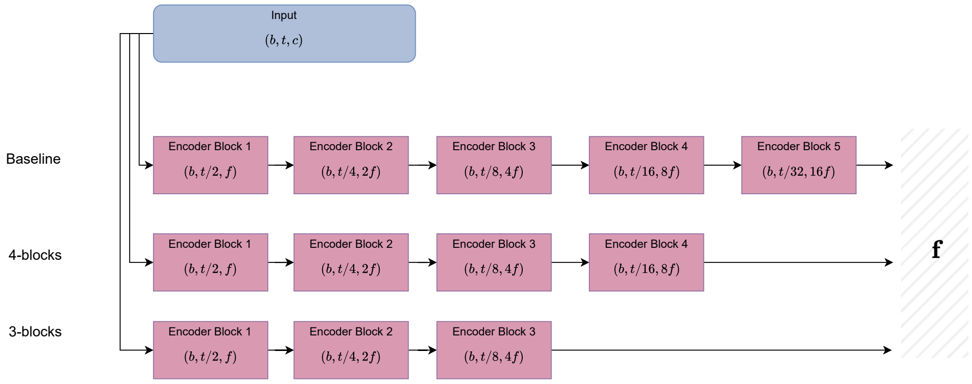


Figure 1. Encoder schematics: the first row shows the baseline model [Remus and da Silveira 2024], and the second and third rows represent the models with four and three ED blocks. Each block is named for convenience and shows the shape of its output in parentheses. The feature map output ($\mathbf{f}$) depends on the number of segments or batches (b), signal length (t), and filters (f). The solid lines represent the layers ordering.

duction through three architectural modifications: (i) reduction of the decoder by merging the two branches into one, (ii) reducing the number of encoder and decoder blocks, and (iii) replacing transposed convolution layers with a parameter-free, interpolation-based upsampling method. This systematic approach leads to five RoI-based model variants for comparative analysis: 5-block ED with interpolation layer (5IL), 4-block ED with transpose convolution (4TC), 4-block ED with interpolation layer (4IL), 3-block ED with transpose convolution (3TC), and 3-block ED with interpolation layer (3IL). For fair comparisons, the same training protocol used by [Remus and da Silveira 2024] was considered here, maintaining the loss function and its hyperparameters Data splitting and RoI definitions are also kept as the baseline method.

Our proposed architectures comprise one encoder ($\mathcal{E}(\cdot)$) and one decoder ($\mathcal{D}(\cdot)$) containing, aside from the shared decoder blocks, two specializing convolution layers: $\mathcal{C}_m(\cdot)$ for the RoI-based fQRS detection and $\mathcal{C}_s(\cdot)$ for fECG recovery. Fig. 1 overviews the baseline and the two options with four and three encoder blocks. Fig. 2 shows the difference between decoders in the baseline and the proposed merged branch, again with five decoder blocks for the baseline and four and three for the reduced ones. We first input the stacked abdominal signals ($\mathbf{a}$) into the encoder ($\mathcal{E}(\cdot)$). The multilevel features extracted by the encoder ($\mathbf{f} = \mathcal{E}(\mathbf{a})$) feed the decoder ($\mathcal{D}(\cdot)$). The last two convolution layers of the decoder are the outputs corresponding to the predicted fECG $\bar{\mathbf{s}} = \mathcal{C}_s(b, t, 32)$ and fQRS mask $\bar{\mathbf{m}} = \mathcal{C}_m(b, t, 32)$, where b and t are the batch size and signal length, and 32 is the number of input features coming from the last decoder block.

The encoder blocks' definition is the same as in the baseline model, and the decoder blocks that use transposed convolution share the same structure as the baseline model. Conversely, when considering interpolation-based decoder blocks, each decoder block has the convolution and transposed convolution layers replaced by an interpolation layer (named `Upsampling1D` in Tensorflow[2]).

[2] `https://www.tensorflow.org/api_docs/python/tf/keras/layers/UpSampling1D`

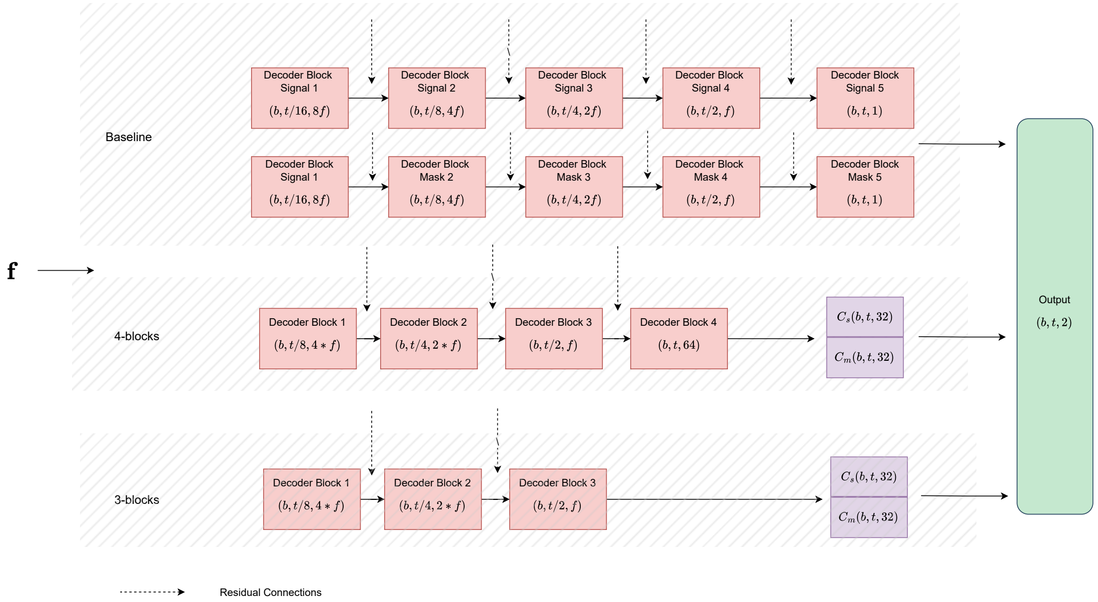

Baseline
Decoder Block Signal 1 $(b, t/16, 8f)$
Decoder Block Signal 2 $(b, t/8, 4f)$
Decoder Block Signal 3 $(b, t/4, 2f)$
Decoder Block Signal 4 $(b, t/2, f)$
Decoder Block Signal 5 $(b, t, 1)$
Decoder Block Signal 1 $(b, t/16, 8f)$
Decoder Block Mask 2 $(b, t/8, 4f)$
Decoder Block Mask 3 $(b, t/4, 2f)$
Decoder Block Mask 4 $(b, t/2, f)$
Decoder Block Mask 5 $(b, t, 1)$
f
4-blocks
Decoder Block 1 $(b, t/8, 4 * f)$
Decoder Block 2 $(b, t/4, 2 * f)$
Decoder Block 3 $(b, t/2, f)$
Decoder Block 4 $(b, t, 64)$
$C_s(b, t, 32)$
$C_m(b, t, 32)$
3-blocks
Decoder Block 1 $(b, t/8, 4 * f)$
Decoder Block 2 $(b, t/4, 2 * f)$
Decoder Block 3 $(b, t/2, f)$
$C_s(b, t, 32)$
$C_m(b, t, 32)$
Output $(b, t, 2)$
Residual Connections

Figure 2. Decoder schematics: the first row shows the baseline model [Remus and da Silveira 2024], and the second and third rows represent the models with four and three ED blocks – convolutional or interpolation-based decoders share the same structure, differentiating only on the definition of the decoder block. Each block is named for convenience and shows the shape of its output in parentheses. The feature map output (f) depends on the number of segments or batches (b), signal length (t), and filters (f). Dashed lines represent the skip connections, and the solid ones represent the layers ordering.

As the baseline, we suppress the batch normalization layer in the convolutional block due to the non-stochastic nature of the ECG signals. We also added dropout layers to avoid overfitting and improve robustness to noise. We keep the first to penultimate residual connections linking the encoder and the decoder blocks. The decoder has the same number of blocks as the encoder. The weights of the decoder blocks are split and feed convolutional layers that specialize the network outputs (signal and mask).

4. Results and Discussion

This section unfolds four, where we present our experimental setup, including the adopted datasets and preprocessing steps, a discussion of the results for the two target tasks, and a thorough comparison with state-of-the-art approaches, including the baseline method.

4.1. Experimental Setup

As we consider supervised deep models, we require ground-truth annotation for training. The annotations provided in real-life datasets [Matonia et al. 2020, Sober and Marco 2007] correspond to the fetal R-peaks and the *invasive* fECG signal (sECG), which is only available during labor where one electrode captures the signal from the fetus scalp; thus, absent in prenatal datasets [Jezewski et al. 2012a]. As in the baseline RoI-based model [Remus and da Silveira 2024], we use the sECG signal stacked with the generated RoI as the ground-truth signals.

We consider two datasets in this study. The Extended ADFECG (XADFECG) dataset [Matonia et al. 2020] unfolds in two: a subset containing antepartum signals of 10 subjects with 20 minutes sampled in 500Hz and a subset of intrapartum signals. In the latter, aECG signals are sampled at 500Hz, and direct fECG signals are sampled at 1000Hz. Five-minute-long recordings for 12 subjects are available. The Non-Invasive Fetal ECG (NI-FECG) dataset [Sober and Marco 2007] contains signals associated with the whole period of one pregnancy. More precisely, the NI-FECG dataset provides two thoracic electrodes and three/four abdominal electrodes. Varying length signals are recorded at 1kHz, and QRS complex annotations are provided.

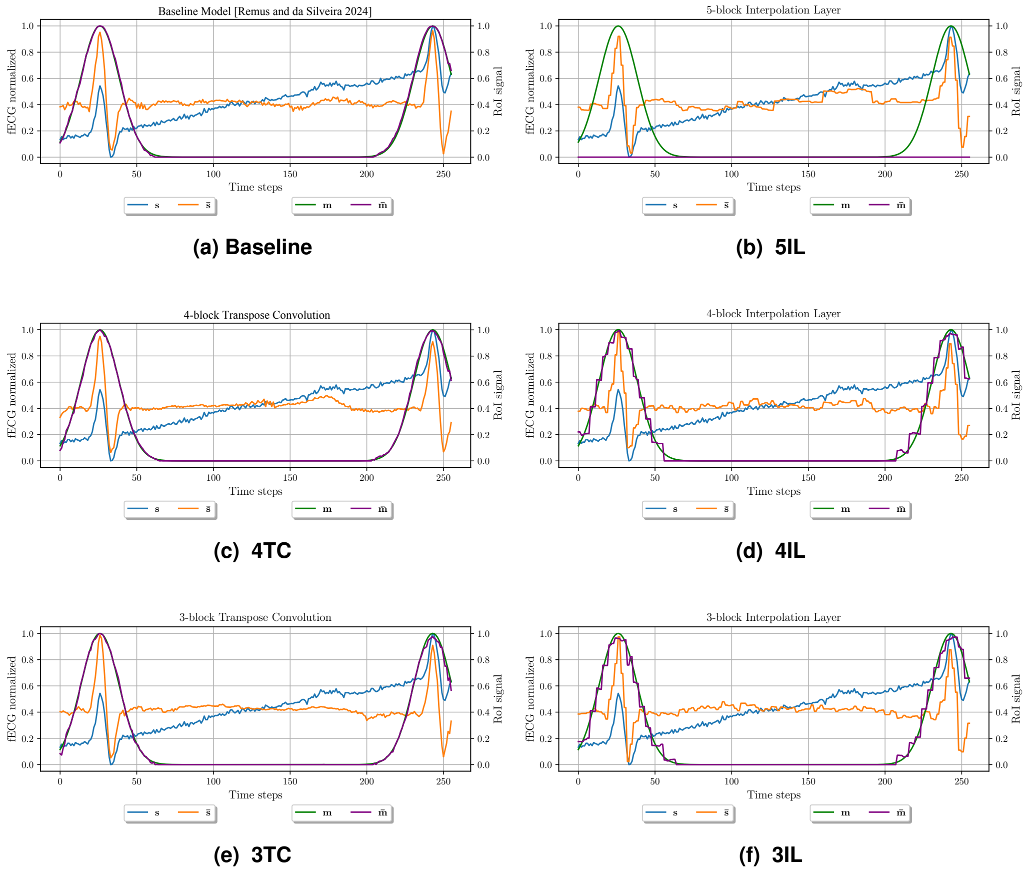
To evaluate the quality and robustness of model training, we applied cross-validation with the leave-one-subject-out method in the XADFECG dataset [Matonia et al. 2020]. Also, cross-dataset inference results in the NI-FECG dataset [Sober and Marco 2007] are provided. As the aECG signals from the XADFECG dataset are sampled at 500Hz, we resampled all other signals to this sampling frequency. The confidence interval (CI) for the results significance was set to 95%. We ran the experiments in a computer equipped with an Intel i7 CPU with 32GB of RAM and an NVIDIA RTX 4070 GPU with 12GB of VRAM.

4.2. fECG Recovery

The recovered fECG comes from the specialized decoder convolution layer $\mathcal{C}_s(\cdot)$. To measure the model's response in this task, we use L_1-norm error metric. We train on the XADFECG dataset and assess on the XADFECG and NI-FECG datasets. For the former, we use a leave-one-subject-out method. We consider ground truth sECG signals extracted via blind source separation for the NI-FECG dataset as in [Remus and da Silveira 2024].

Fig. 3 shows qualitative results for each model trained for one specific segment on `r11` subject from the XADFECG dataset. On the left column, we show the models based on transposed convolution layers; on the right, the models based on interpolation layers. The rows show the number of encoder and decoder blocks used, starting with five (baseline). Overall, the model can reconstruct the fECG, ignoring environmental noises from the raw sECG. In comparing decoder methodologies, we note that using the transposed convolution layer provides a smoother reconstruction for RoI masks and fECG signals. However, the R-peak location is well defined in all cases, excluding 5IL, discussed in the following section, highlighting the feature extraction characteristic from the encoder. In the TC-based models, it can be seen that models with fewer ED blocks are less susceptible to reconstructing high-frequency noises, though the low-frequency noise is well ignored in both cases. The reconstruction of low-frequency noises is more prominent in the 4TC model. Although the 5IL model can reconstruct the signal well (see Fig. 3b), the linear baseline is not filtered out compared with the other models.

The average reconstruction $L1$-error for the XADFECG and NI-FECG datasets for each model is presented in Table 1. As observed, there is not much difference in reconstruction errors between the models in terms of the average and the confidence interval (CI) range for both datasets. The 3-block ED model exhibits the lowest average error, which can be attributed to its greater similarity to the sECG signal rather than a filtered fECG. Additionally, this model is less noise-sensitive and produces a smoother signal due to having fewer parameters. In the NI-FECG dataset, inference is performed using a



(a) Baseline

(b) 5IL

(c) 4TC

(d) 4IL

(e) 3TC

(f) 3IL

Figure 3. Example of fECG/RoI mask reconstruction for subject `r11` in XADFECG dataset, trained with XADEFCG dataset in leave-one-subject-out method. In blue and in green are the ground-truth signals, respectively sECG (s) and RoI (m), and in purple and orange, the predicted signals: fECG ($\bar{\mathrm{s}}$) and RoI ($\bar{\mathrm{m}}$). The models with transpose convolution layers are shown on the left column, and on the right, the models are based on interpolation layers. The rows show the number of encoder and decoder blocks used, starting with five. The top-left corresponds to the baseline model.

pre-trained model, meaning results may vary across different inference runs. The values presented represent the average across the evaluation files. The higher error observed in the NI-FECG case can be attributed to the comparison with a retrieved signal rather than the original sECG, which is not provided. Future research should involve healthcare professionals to improve signal interpretation and reconstruction, since the available datasets only provide sECG data for reconstruction comparison.

4.3. fQRS Detection

The RoI mask comes from the specialized decoder convolution layer $\mathcal{C}_m(\cdot)$. We use the center of the estimated Gaussian(s) as the R-peak position(s) as our fQRS detector. We use a regular peak finder[3] with a minimum threshold of 0.7 and a minimum peak distance of 0.2 seconds. Note that we feed it with the recovered RoI masks instead of applying a peak finder on the recovered fECG as [Zhong et al. 2019, Mohebbian et al. 2022,

[3] We consider the algorithm from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html`.

Table 1. Average fECG signal reconstruction $L1$-error. Results are shown for the XADFECG cross-validation method and NI-FECG inference with different models trained on the XADFECG dataset. Base is the baseline [Remus and da Silveira 2024], the suffix *IL* is for models with decoder based on interpolation layers while *TC* is for those based on transposed convolution; the prefix number represents the number of ED pairs.

Dataset	Model					
	Base	5IL	4TC	4IL	3TC	3IL
XADFECG	0.1418 ± 0.0117 (0.1189 - 0.1646)	0.1435 ± 0.0098 (0.1243 - 0.1628)	0.1404 ± 0.0117 (0.1174 - 0.1634)	0.1391 ± 0.0109 (0.1177 - 0.1604)	0.1368 ± 0.0113 (0.1146 - 0.1589)	0.1388 ± 0.0111 (0.117 - 0.1606)
NI-FECG	0.2949 ± 0.0191 (0.2575 - 0.3323)	0.2678 ± 0.0199 (0.2287 - 0.3069)	0.2921 ± 0.0242 (0.2446 - 0.3395)	0.2925 ± 0.0173 (0.2586 - 0.3264)	0.2976 ± 0.0149 (0.2685 - 0.3268)	0.2814 ± 0.015 (0.252 - 0.3109)

Ghonchi and Abolghasemi 2022, Rahman et al. 2023].

The considered metrics for the fQRS detection task are recall, precision, and F1-score, as in [Zhong et al. 2019, Mohebbian et al. 2022, Ghonchi and Abolghasemi 2022, Rahman et al. 2023, Remus and da Silveira 2024], providing the mean and standard deviation with a CI of 95%. Table 2 shows the results per subject on the XADFECG dataset (the subject left for inference given the model trained with the others). Table 3 presents the results per file (pregnancy time) for each trained model with the XADFECG dataset.

From Table 2, we note that all models perform poorly when testing on subject `r03`. Matonia et al. [Matonia et al. 2020] explain two significant problems in the recordings for that subject: a 27.5% signal loss and a negative $\log$ relation (WF) between the fECG signal and the aECG. Another highlighted information is that on 5IL model, the inference on `r10` and `r11` return null information as the RoI mask is not reconstructed, as can be seen in Fig. 3b. In these specific cases, we hypothesize that the training achieved the minimum loss value while ignoring the reconstruction of the RoI mask; as it can be seen on loss equation [Remus and da Silveira 2024], the third element combines the predicted mask and ground-true signal and ground-true mask and predicted signal, leaving space for the minimization without the good reconstruction of the mask.

Still, increasing ED blocks leads to a higher recall in the recall metric. Additionally, models based on transposed convolution demonstrate better recall results than their counterparts using interpolation layers. From this combination of results, we can conclude that models with more parameters exhibit a stronger ability to retrieve true fQRS complex positions, particularly in noisier measurements with $WF \leq 0$ (`r03` and `r07`). Meanwhile, the precision metric does not appear to be influenced by the number of ED pairs, though models with interpolation layers show slightly better precision. We hypothesize that transpose convolution-based models, having more parameters, are more sensitive to signal variations—beneficial for recall but detrimental to precision. A potential way to enhance the robustness of convolution-based models is to train them with data exhibiting more significant variability. Overall, the F1-score metric indicates that, on average, decoders with 5 and 4 blocks of transpose convolutions outperform the other approaches.

NI-FECG results are presented in Table 3 for each model proposed per file inferred. The file numbers are in ascending order based on the pregnancy time measured. To make a fairer comparison, the files evaluated were the same as in [Mohebbian et al. 2022] and [Remus and da Silveira 2024]. The precision metric evolution along the pregnancy

Table 2. FQRS complex detection results per subject on the leave-one-subject-out cross-validation on the XADFECG dataset. Acronyms are the same as in Table 1.

Metric (%)	Model	r01	r02	r03	r04	r05	r06	r07	r08	r09	r10	r11	r12	Mean ± std. dev. (CI 95%)
Precision	Base	99.38	95.28	74.68	96.62	99.69	99.85	97.65	99.22	98.20	97.80	98.31	97.70	96.2 ± 1.99 (92.29 - 100)
	5IL	99.38	95.52	79.43	97.74	100	100	98.23	99.38	98.78	0	0	98.89	80.61 ± 10.99 (59.07 - 100)
	4TC	99.53	94.40	77.40	96.0	99.54	99.41	97.44	98.76	96.10	99.51	98.17	97.49	95.12 ± 2.62 (89.98 - 100)
	4IL	99.37	94.54	77.14	97.98	99.23	99.25	97.85	98.44	96.22	99.50	97.85	96.38	96.15 ± 1.78 (92.66 - 99.64)
	3TC	99.84	93.41	73.05	96.68	100	98.67	97.14	98.16	96.98	99.01	98.31	95.33	95.55 ± 2.12 (91.4 - 99.7)
	3IL	99.84	93.57	78.73	97.53	99.84	99.55	95.11	98.12	98.00	99.50	98.16	96.65	96.27 ± 1.69 (92.95 - 99.59
Recall	Base	99.22	98.58	66.11	97.77	99.09	99.71	98.73	98.91	97.03	99.36	99.38	97.26	95.85 ± 2.72 (90.52 - 100)
	5IL	99.69	97.48	62.18	95.59	96.96	98.39	96.99	99.22	96.28	0	0	94.97	78.15 ± 10.94 (56.71 - 99.58)
	4TC	99.22	98.28	58.54	95.29	98.94	99.12	96.35	99.38	95.39	97.28	99.53	94.66	94.33 ± 3.29 (87.88 - 100)
	4IL	99.53	98.26	52.94	93.09	97.87	97.80	93.98	98.60	94.65	95.37	98.91	93.29	92.86 ± 3.69 (85.63 - 100)
	3TC	99.69	98.27	34.17	94.26	97.57	98.53	96.99	99.37	95.54	96.64	99.54	96.49	92.26 ± 5.3 (81.87 - 100)
	3IL	99.07	98.58	34.73	92.94	96.35	98.39	95.72	98.91	95.09	95.53	99.22	96.04	91.77 ± 5.22 (81.54 - 100
F1-score	Base	99.29	96.90	70.13	96.69	99.39	99.78	98.19	99.07	97.61	98.57	98.84	97.48	96.0 ± 2.37 (91.35 - 100)
	5IL	99.53	96.49	69.76	96.65	98.46	99.19	97.61	99.30	97.52	0	0	96.89	79.28 ± 10.95 (57.83 - 100)
	4TC	99.38	96.30	66.67	95.65	99.23	99.27	96.89	99.07	95.75	98.38	98.84	96.06	96.15 ± 1.77 (92.68 - 99.62)
	4IL	99.46	96.37	62.79	95.46	98.55	98.52	95.88	98.52	95.43	97.39	98.38	94.81	94.3 ± 2.9 (88.62 - 99.98)
	3TC	99.77	95.78	46.56	95.46	98.77	98.60	97.06	98.76	96.25	97.82	98.92	95.91	93.31 ± 4.27 (84.94 - 100)
	3IL	99.45	96.01	48.20	98.18	98.07	98.97	95.41	99.22	96.53	97.47	98.69	96.26	93.29 ± 4.12 (85.21 - 100)

Table 3. FQRS complex detection results per file inferred on the NI-FECG dataset. The file numbers are in ascending order based on the pregnancy time measured. Acronyms are the same as in Table 1.

Metric (%)	Model	154	192	811	274	323	368	826	244	290	733	597	746	906	444	Mean
Precision	Base	75.20	74.46	75.74	87.97	81.66	82.69	74.33	87.30	92.92	92.83	94.09	88.43	95.59	94.21	85.53 ± 2.16 (81.3 - 89.76)
	5IL	76.81	75.47	77.74	94.3	83.12	82.77	74.41	89.48	92.97	92.97	94.43	88.13	96.01	94.36	86.64 ± 2.14 (82.44 - 90.84)
	4TC	84.45	75.55	82.25	97.22	76.02	83.13	73.36	93.26	94.86	94.11	95.01	88.75	96.46	95.06	87.82 ± 2.28 (83.34 - 92.3)
	4IL	75.28	73.11	68.82	94.01	68.83	77.51	64.02	87.49	89.98	90.32	91.89	83.14	94.33	91.79	82.18 ± 2.84 (76.61 - 87.75)
	3TC	86.43	81.89	75.88	97.12	79.95	83.48	72.34	95.35	95.72	96.59	96.39	93.99	97.35	97.02	89.25 ± 2.39 (84.56 - 93.94)
	3IL	96.07	93.97	92.87	99.52	90.36	91.49	86.51	98.27	98.79	98.32	98.55	97.95	99.20	98.90	95.77 ± 1.09 (93.64 - 97.9)
Recall	Base	99.79	97.82	100	100	100	100	100	100	100	100	99.96	100	98.28	100	99.7 ± 0.19 (99.33 - 100)
	5IL	99.79	96.51	99.59	100	100	99.85	99.76	100	100	100	99.83	99.60	95.43	99.91	99.3 ± 0.38 (98.55 - 100)
	4TC	99.79	96.95	99.59	100.0	100.0	99.56	99.76	100.0	99.89	99.83	99.87	99.43	96.0	99.72	99.31 ± 0.33 (98.67 - 99.96)
	4IL	99.57	97.17	98.97	99.84	99.54	99.41	99.28	99.71	99.68	99.66	99.31	98.79	94.45	99.50	98.92 ± 0.39 (98.16 - 99.68)
	3TC	99.36	96.51	71.49	96.50	98.77	71.01	68.84	69.17	63.54	64.32	70.00	68.17	52.55	70.26	75.75 ± 4.07 (67.77 - 83.73)
	3IL	99.36	94.99	86.16	99.52	99.54	80.23	78.99	76.78	65.58	70.00	70.05	73.34	43.45	70.35	79.17 ± 4.25 (70.84 - 87.49)
F1-score	Base	85.77	84.56	86.20	93.60	89.90	90.52	85.27	93.22	96.32	96.28	96.94	93.86	96.92	97.02	91.88 ± 1.28 (89.38 - 94.39)
	5IL	86.8	84.7	87.32	97.07	90.78	90.51	85.24	94.45	96.36	96.36	97.05	93.52	95.72	94.45	92.35 ± 1.26 (89.89 - 94.81)
	4TC	91.48	84.92	90.09	98.59	86.38	90.61	84.54	96.51	97.31	96.88	97.38	93.79	96.23	97.33	93.00 ± 1.33 (90.39 - 95.62)
	4IL	85.74	83.44	81.19	96.84	81.38	87.11	77.84	93.20	94.58	94.76	95.46	90.29	94.39	95.49	89.41 ± 1.73 (86.02 - 92.8)
	3TC	92.45	88.60	73.62	96.81	88.37	76.74	70.54	80.18	76.38	77.22	81.11	79.03	68.26	81.50	80.77 ± 2.2 (76.47 - 85.07)
	3IL	97.69	94.47	89.39	99.52	94.73	85.49	82.58	86.21	78.83	81.78	81.89	83.88	60.43	82.22	85.65 ± 2.62 (80.52 - 90.78)

interval shows that the models based on 5 and 4 ED pairs have an improvement along the pregnancy time; in opposition, the 3-block models seem not to be affected by the pregnancy week. One possible reason is that the fetal signal is weaker in the earlier weeks and has a bigger fHR compared to measurements closer to labor. Thus, as the segments are normalized and have a length of 0.5s, it is possible to have segments without present fQRS complexes; in this case, noise is amplified and captured as fQRS complexes by the model. Meanwhile, the 3-block EDs have fewer parameters and are more robust to noise. However, as one may notice, in the recall metric, the 3-block ED models present worse results during the pregnancy evolution, showing that the models did not learn the correct pattern to find the fQRS complexes in segments. 5- and 4-block EDs maintain good results, larger than 90% of recall, in all analyzed files. Overall, the F1-metric results show that the 4-block transpose convolution model (4TC) is the better option.

4.4. Comparison with the State-of-the-Art

There is no standardized benchmarking for fECG recovery, as studies employ different training datasets, preprocessing steps, and input data formats, making comparisons challenging. Plus, not all studies assess the morphological quality of fECG extraction. Fortunately, all reviewed works report fQRS complex detection metrics.

Authors from [Rahman et al. 2023], [Mohebbian et al. 2022], and [Wang et al. 2023] reported F1-scores of 99.6% ± 0.2% (no CI provided), 99.7% ± 0.4% (97.8 - 99.9), and 99.71% ± 0.10% (no CI provided), respectively, for fQRS complex detection using the ADFECG dataset. Zhong et al. [Zhong et al. 2019] obtained an F1-score of 94.10% ± 0.6422% (92.83 - 95.36) on the same dataset but used a different training protocol. Additionally, Ghonchi et al. [Ghonchi and Abolghasemi 2022] achieved 92.87% on the NI-FECG dataset, while Mohebbian [Mohebbian et al. 2022] reported 97.9% when training with the ADFECG dataset. On the XADFECG dataset, Rahman et al. [Rahman et al. 2023] achieved 99.5% ± 0.3%, and Wang et al. [Wang et al. 2023] reported 99.47% ± 0.17%, with no confidence intervals provided for either model. The baseline model [Remus and da Silveira 2024] achieved results comparable to previous methodologies, with an F1-score of 96.0 ± 2.37 (91.35 - 100) on the XADFECG dataset and 91.88 ± 1.28 (89.38 - 94.39) for NI-FECG dataset. The baseline model simultaneously reconstructs the fECG signal and detects fQRS locations while incorporating a noise-aware protocol.

Regarding the proposed models, as shown in Table 2 and Table 3, the 4TC model appears to yield better results, possibly due to its smaller number of parameters, making it less susceptible to noise and more generalizable than the baseline. The 5IL model also has competitive performance; however, since it failed to return any RoI masks for subjects `r10` and `r11`, this could affect other inferences, although the final model is not trained using the leave-one-subject-out method. The 4IL model exhibits a size reduction and inference time comparable to the 4TC version, but its fECG signal reconstruction lacks the smoothness of convolutional models, making it a less favorable option. Models based on three-block architectures perform worse across both datasets, mainly due to a decline in recall as pregnancy progresses, which impairs their usability. However, if this issue is addressed, these models could still be viable. Despite having the fewest parameters among the proposed models, their inference time does not improve proportionally.

Furthermore, the inference times per fECG minute recorded by our methods and the baseline are lower than [Mohebbian et al. 2022], with 1.9 ± 0.4s, [Ghonchi and Abolghasemi 2022], with 1.32s and [Wang et al. 2023], with 0.97s (considering that the time reported in their work is for one 5 minute long dataset file).

5. Conclusion

This work assesses different approaches for reducing the size of a state-of-the-art method for fECG recovery and fQRS complex detection [Remus and da Silveira 2024], opening room for implementing such simplified approaches in low-cost battery-powered equipment. Five alternatives are compared with the baseline: replacing transpose convolution with an interpolation layer and reducing to 4 and 3 ED blocks in the architecture. We considered inter-subject and cross-dataset experiments considering the XADECG and NI-FECG datasets. The results show that a model with 4 ED blocks and transposed convolution is a promising alternative to the baseline, providing a five-fold reduction in the model size. For use in real applications, future work can consider quantizing the models to facilitate hardware implementation, evaluating them on more diverse datasets, and in-

Table 4. Model size and inference time comparison between models. NI stands for not informed in the article. The number of parameters in the previous works was retrieved by implementing the models. Acronyms are the same as in Table 1.

Model	Number of parameters	Size (MB)	Inference time per subject per recorded ECG minute (s)
Base	21,010,242	80.15	0.066
5IL	15,790,850	60.24	0.052
4TC	4,631,234	17.67	0.034
4IL	4,003,522	15.27	0.033
3TC	1,152,450	4.40	0.032
3IL	984,002	3.75	0.028
[Zhong et al. 2019]	113,617	0.44	NI
[Mohebbian et al. 2022]	8,756	0.033	1.9
[Ghonchi and Abolghasemi 2022]	5,674,241	21.65	1.32
[Rahman et al. 2023]	6,160,225	23.50	NI
[Wang et al. 2023]	25,235,136	96.28	0.97

volving a multidisciplinary team to assess how to improve and adapt the tool for practical scenarios.

Acknowledgments

The authors thank the *Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul* for partially funding this research.

References

Andreotti, F. et al. (2016). An open-source framework for stress-testing non-invasive foetal ECG extraction algorithms. *Physiological Measurement*, 37(5):627–648.

Barnova, K. et al. (2024). Artificial intelligence and machine learning in electronic fetal monitoring. *Archives of Computational Methods in Engineering*.

Behar, J. et al. (2016). A practical guide to non-invasive foetal electrocardiogram extraction and analysis. *Physiological Measurement*, 37(5):R1–R35.

Cohen, R. et al. (2024). Looks too good to be true: An information-theoretic analysis of hallucinations in generative restoration models. *arXiv preprint arXiv:2405.16475*.

Ghonchi, H. and Abolghasemi, V. (2022). A dual attention-based autoencoder model for fetal ECG extraction from abdominal signals. *IEEE Sensors Journal*, 22(23):22908–22918.

Jezewski, J. et al. (2012a). Determination of fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. *Biomedizinische Technik Biomedical Engineering*, 57(5).

Jezewski, J. et al. (2012b). Determination of fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. *Biomedizinische Technik/Biomedical Engineering*, 57(5).

Kahankova, R. et al. (2020). A review of signal processing techniques for non-invasive fetal electrocardiography. *IEEE Reviews in Biomedical Engineering*, 13:51–73.

Liu, Y. et al. (2025). Research on approximate computation of signal processing algorithms for aiot processors based on deep learning. *Electronics*, 14(6):1064.

Matonia, A. et al. (2020). Fetal electrocardiograms, direct and abdominal with reference heartbeat annotations. *Scientific Data*, 7(1).

Mendis, L. et al. (2023). Computerised cardiotocography analysis for the automated detection of fetal compromise during labour: A review. *Bioengineering*, 10(9):1007.

Mohebbian, M. R. et al. (2022). Fetal ECG extraction from maternal ECG using attention-based cyclegan. *IEEE Journal of Biomedical and Health Informatics*, 26(2):515–526.

Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236.

Rahman, A. et al. (2023). Fetal ecg extraction from maternal ecg using deeply supervised LinkNet++ model. *Engineering Applications of Artificial Intelligence*, 123:106414.

Remus, J. C. and da Silveira, T. L. T. (2024). An end-to-end roi-based encoder-decoder for fetal ecg recovery and qrs complex detection. In *International Symposium on Medical Measurements and Applications*, volume 40, page 1–6. IEEE.

Sameni, R. (2021). Noninvasive fetal electrocardiography: Models, technologies, and algorithms. In *Innovative Technologies and Signal Processing in Perinatal Medicine: Volume 1*, volume 1, chapter 5. Springer International Publishing.

Shi, X. et al. (2023). Unsupervised learning-based non-invasive fetal ECG muti-level signal quality assessment. *Bioengineering*, 10(1):66.

Shuvo, M. M. H. et al. (2023). Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proceedings of the IEEE*, 111(1):42–91.

Silva, I. et al. (2013). Noninvasive fetal ECG: the physionet/computing in cardiology challenge 2013. *Computing in cardiology*, 40:149–152.

Sober, M. M. and Marco, J. G. (2007). Non-invasive fetal electrocardiogram database.

UNICEF (2019). Healthy mothers, healthy babies: Taking stock of maternal health. Technical report, United Nations Children's Fund.

Vullings, R. et al. (2009). Dynamic segmentation and linear prediction for maternal ecg removal in antenatal abdominal recordings. *Physiological Measurement*, 30(3):291–307.

Wang, X. et al. (2023). Correlation-aware attention cyclegan for accurate fetal ecg extraction. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13.

WHO, editor. *Trends in maternal mortality 2000 to 2020.* World Health Organization, Geneva.

Zhong, W. et al. (2019). Fetal electrocardiography extraction with residual convolutional encoder–decoder networks. *Australasian Physical amp; Engineering Sciences in Medicine*, 42(4):1081–1089.