LSTM-only Model for Low-complexity HR Estimation from Wrist PPG

Leandro Giacomini Rocha[1], Guilherme Paim,[2] Dwaipayan Biswas[3], Sergio Bampi[2], Francky Catthoor[4], Chris Van Hoof[3], Nick Van Helleputte[3]

***Abstract*— Continuous and non-invasive cardiovascular monitoring has gained attention due to the miniaturization of wearable devices. Particularly, wrist-worn photoplethysmography (PPG) sensors present an alternative to electrocardiogram recording for heart rate (HR) monitoring as it is cheaper and non-intrusive for daily activities. Yet, the accuracy of PPG measurements is heavily affected by motion artifacts which are inherent to ambulatory environments. In this paper, we propose a low-complexity LSTM-only neural network for HR estimation from a single PPG channel during intense physical activity. This work explored the trade-off between model complexity and accuracy by exploring different model dataflows, number of layers, and number of training epochs to capture the intrinsic time-dependency between PPG samples. The best model achieves a mean absolute error of 4.47 ± 3.68 bpm when evaluated on 12 IEEE SPC subjects.**

***Clinical relevance*– This work aims to improve the quality of HR inference from PPG signals using neural network, enabling continuous vital signal monitoring with little interference in daily activities from embedded monitoring devices.**

I. Introduction

Heart rate (HR) estimation is a key element of wearable health monitoring applications. Photoplethysmography (PPG) sensors have been widely adopted in the commercial development of wrist-worn sensors as they can be placed in-body extremities without interfering with the subject's daily activities. These PPG signals rely on pulse oximeters composed of a light-emitting diode and a photodetector(PD) that may operate either in transmittance or reflection mode. The PD detects the variations of the intensity of the transmitted/reflected light to capture the cardiac rhythm [1].

Despite the low cost and pervasiveness of such devices, they are highly susceptible to motion artifacts (MA), significantly reducing the reliability of vital parameters measurements [2]. Several components contribute to signal corruption by MA, like the separation between the sensor from the skin surface and the body movements that leads to changes in blood flow, overpowering the heartbeat spectrum. Fostered by the IEEE Signal Processing Cup (SPC) 2015 [3], MA attenuation from wrist-worn PPG sensors has been explored using traditional digital signal processing and machine learning techniques. State-of-the-art works achieved satisfactory accuracy on the SPC dataset through complex models [4].

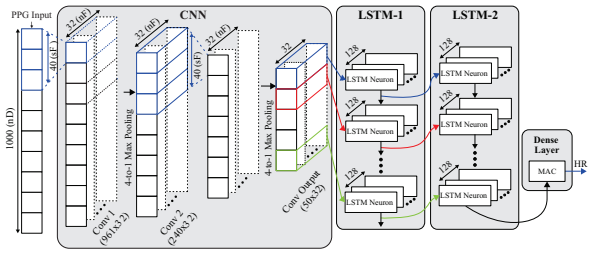

Fig. 1: Motivation: state-of-the-art DNN architecture for HR estimation from wrist PPG combines CNN and LSTM layers [5]–[7].

The authors of [1] proposed a deep neural network (DNN) that combines a convolutional neural network (CNN)-based feature extractor along with two long short-term memory (LSTM) layers and a dense layer to capture the temporal dependencies on a single PPG channel captured from a wrist-worn device. Despite the improvements proposed in [5], [6] (Fig. 1), this DNN is highly complex with more than 250k parameters and 20 million multiply-accumulate (MAC) operations per inferred HR. In that sense, we propose an exploratory study aiming for an LSTM-only network model for low-complexity HR estimation from PPG signals maintaining similar accuracy to the state-of-the-art. Our idea relies on the ability of recurrent neural network-based models to estimate output based on past inputs sequences and the current hidden state for either univariate or multivariate series. This can be achieved by modeling the sampled PPG signal as a univariate time series, which can predict an HR. We explore the model structure and hyperparameters with a grid search to improve the generalization capability of the model and its accuracy. We evaluate our LSTM-only network proposal on 12 IEEE SPC subjects achieving a mean average error (MAE) of 4.47 ± 3.68 bpm on the best network model. Hence, the **key contribution of our paper** is the investigation of LSTM-only network architectures for reduced complexity, i.e., fewer operations per inference.

II. Motivation and Problem Formulation

Current research on wrist PPG has mostly focused on mitigating the effects of motion artifacts on PPG sensors with green LED due to their short wavelength compared to other light sources [3]. Deep learning models are a promising strategy to estimate HR in these applications as they achieve comparable performance to traditional digital signal processing (DSP) algorithms with fewer sensor signals [7]. In this context, recurrent neural networks (RNN) are an effective approach to process time-series data like PPG signals for HR estimation. The key feature of these models is the ability to

[1]L. G. Rocha is with the Federal Institute of Education, Science and Technology, Sertão, Brazil; e-mail: leandro.rocha@sertao.ifrs.edu.br

[2]G. Paim and S. Bampi are with the Graduate Program in Microelectronics (PGMicro), Informatics Institute, Federal University of Rio Grande do Sul, Porto Alegre, Brazil; e-mail: {gppaim, bampi}@inf.ufrgs.br

[3]D. Biswas, C. V. Hoof, N. V. Helleputte are imec, Kapeldreef 75, 3001 Leuven, Belgium; e-mail: {Dwaipayan.Biswas, Chris.VanHoof, Nick.VanHelleputte}@imec.be

[4]F. Catthoor is with ESAT Department, KU Leuven and imec, Leuven, Belgium,; e-mail: Francky.Catthoor@imec.be

observe and integrate contextual information from previous inputs and combine them with current inputs, improving the model's robustness to time distortions that may occur on the input sequence. Nonetheless, vanilla RNNs suffer from the vanishing gradient when long input sequences are considered [8]. This issue was addressed with the proposal of LSTM networks, which have internal memory cells and a couple of adaptive and multiplicative gating units on the input and output of all cells. These mechanisms control the information flow in each timestep, effectively constraining the influence on previous hidden states on the current predictions. These gates, known as input (i_t), forget (f_t), output (o_t), and cell state update (u_t), and they compute the intermediary results from the current input (x_t) and the previous hidden state (h_{t-1}) as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \tag{2}$$

$$o_t = \sigma(W_o x_o + U_o h_{t-1}) \tag{3}$$

$$u_t = \tanh(W_u x_t + U_u h_{t-1}) \tag{4}$$

$$c_t = f_t \times c_{t-1} + i_t \times u_t \tag{5}$$

$$h_t = o_t \times \tanh(c_t) \tag{6}$$

In these equations, x_t is the LSTM input, W_* and U_* are weight matrices, and h_t is the hidden state, which becomes the input h_{t-1} on the next timestep and c_{t-1} is the cell state from the previous timestep. All bias terms were omitted for simplicity's sake. The σ and $\tanh$ represent the sigmoid and hyperbolic tangent non-linear functions, respectively.

Edge devices often lack computing effort and energy availability in batteries to accomplish these tasks [9]. Hence, optimizations on the software model can be further replicated into custom hardware accelerators to maximize performance with the least amount of energy. In that sense, DL solutions more closely tied to the hardware implementation will be preferred due to the unique hardware/software co-design possibility to boost energy efficiency [10].

Current state-of-the-art HR estimation using deep learning was proposed in [1] where the authors combine convolutional layers along with LSTM layers and a final regression layer, presenting an end-to-end framework for training and evaluation without manual feature extraction. The proposed model has 256k parameters, and it requires around 20 million operations per inferred HR. A reduced complexity implementation of this model was proposed in [6] where binarization and quantization techniques were explored, although the number of operations and parameters remained constant. A similar approach was proposed in [11] where the authors use a single PPG channel to predict the HR and blood pressure simultaneously. They adopt a different windowing scheme for the input data, which reduces the model complexity when compared to [1] even though the model predicts additional physiological data. However, a direct accuracy comparison is not feasible as these works target different datasets.

Our exploration adopts the IEEE SPC database as it is the most employed dataset in the literature to evaluate the algorithm performance. This dataset contains 5-minute records from 22 healthy subjects with ages ranging from 18 to 58 [3]. Each record comprises two PPG channels captured from a wrist-worn oximeter with a green LED along with a 3-channel accelerometer to measure the wrist movements. The records also have an ECG signal captured from a chest-worn patch used as the ground truth for HR estimation. All signals were sampled at 125 Hz, and they were divided into 8s windows with an overlap of 6s with the adjacent windows (sliding windows by 2s). We limited the dataset to the first 12 subjects, which executed a well-defined activity protocol that involved walking and running on a treadmill according to the following protocol: 1–2 km/h for 0.5 min, 6–8 km/h for 1 min, 12–15 km/h for 1 min, 6–8 km/h for 1 min, 12–15 km/h for 1 min, and 1–2 km/h for 0.5 min. Limiting the dataset is necessary due to the long training time and the number of experiments executed to obtain the optimal network structure.

III. LSTM-only Network Proposal for Heart-Rate Estimation from Wrist PPG

Despite all the information channels on the SPC dataset, we adopted an approach similar to [1] and limited the network to operate with a single PPG channel with the ground truth generated by the ECG signal to reduce the model complexity. Hence, we search for the most suitable model by exploring three optimization axis: (i) the dataflow model, (ii) the number of layers, and (iii) the optimal set of hyperparameters for network training. The last two optimization axis will be explored in Section IV.

The LSTM networks were trained on an Nvidia Tesla K80 GPU (with cuDNN 5.10) and modeled in Lasagne 0.2dev1 [12], configured to use Theano 0.9.0 as the backend. The first 12 healthy male subjects of the SPC dataset [3] was used to train and evaluate the network.

As we aim for a more generalized model, we adopt the *leave-one-subject-out* (LOSO) training methodology [6]. This strategy removes the subject under test (SUT) from the dataset and uses the remaining subject data as the training dataset, improving the model capability to capture the physiological variance among subjects. The LOSO method also ensures that there is no overlap between train and test datasets. The models adopt a learning rate of 0.005 with a batch size of 16 samples and a logcosh function as the loss criteria. The accuracy measurement is achieved by computing, for each window, the absolute difference between the ground truth HR and the predicted HR, providing the information to compute the MAE and the standard deviation (SDAE) for all windows.

The first optimization point relies on how the data is fed to the network and how it will predict the output based on current input and state. Therefore, this work explores the LSTM-only networks in the three following dataflow models:
Model I: The first model is the sequence-to-one dataflow where inputs are fed continuously, but HR estimation is only generated at the last time step (see Fig. 2a). Each input is a scalar value (S* on the figure), and the timestep is 1000 as

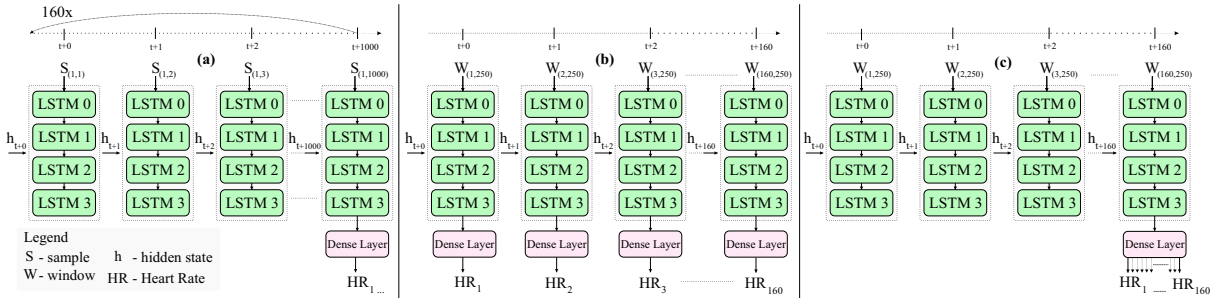

Fig. 2: Explored LSTM-only network dataflows. (a) Sequence-to-one (Model I). (b) Sequence-to-sequence (Model II). (c) Hybrid approach (Model III).

each time window has 1000 samples.

Model II: The second model is based on the sequence-to-sequence dataflow where an output is generated in each timestep (Fig. 2b. Each input is a window (W* on the picture) with 250 samples (equivalent to 2s of data). There is a predicted HR for each input window, updating the hidden state between inputs.

Model III: The third model can be regarded as the combination of the first and second models as the input sequence consists of complete subject information because the input vector is N windows of 250 samples each. We assume that N is 160 as this is the maximum number of windows for a given subject on the dataset. For records with a lower number of windows, the last window is replicated to ensure that all records are of the same size. Once all N windows have been processed, the hidden state of the last LSTM layer is fed to a dense layer with 160 neurons (1 for each window) to execute the linear regression for the HR estimation for each window. As it needs to process all subject's input data before any valid output is processed, this model has the highest latency.

IV. Results and Discussions

This section shows the evaluation results. Firstly, we present the training details. Then, we present the trade-off between the number of LSTM layers versus the accuracy.

The LSTM models are trained using the strategy leave-one-subject-out explained in [6]. LSTMs are particularly sensitive to the number of training epochs. Too many iterations during the training phase could result in over-fitting, and the model is no longer generalized. On the opposite, insufficient training iterations leads to bad regression accuracy. For this exploration, we swept the number of training epochs from 1 to 300 and observed that longer training leads to a model too specialized whose performance is degraded when evaluated on the test dataset. The search found that the optimal training length is 30 epochs as it corresponds to the minimal validation loss as longer runs led to overfitting.

In LSTM models, increasing the number of layers has diminishing returns in terms of accuracy improvement of the model. As more layers are stacked, the model complexity increases, leading to difficulties in the training process. A sweeping analysis was executed evaluating the correlation between the MAE and the number of layers to obtain the optimal model for HR inference.

Fig. 3a illustrates how the model accuracy is affected by the number of stacked layers. As more layers are added, it becomes more difficult to train the parameters, as shown by the training loss. Since the dataset is not large enough, it limits the efficiency of the training algorithm when the number of parameters grows. Nonetheless, increasing the number of layers improves the accuracy by nearly 50%, reducing the MAE to a plateau of 4.3 bpm around four layers in comparison with the 1-layer model, which presented an MAE of 8.0 bpm. Although models with 3 and 4 layers have similar performance, Fig. 3b shows that the 4-layer model has a lower variability on the predicted output, although this approach requires a slightly higher number of parameters.

Table I illustrates how the dataflow influences the model accuracy and compares the proposed model with the current state-of-the-art in HR estimation from PPG data using deep learning algorithms. In all cases, the model evaluation assumes the optimal number of layers/hyperparameters that were obtained through our analysis, which is explained in sub-section B and C. Model III presents the lowest overall

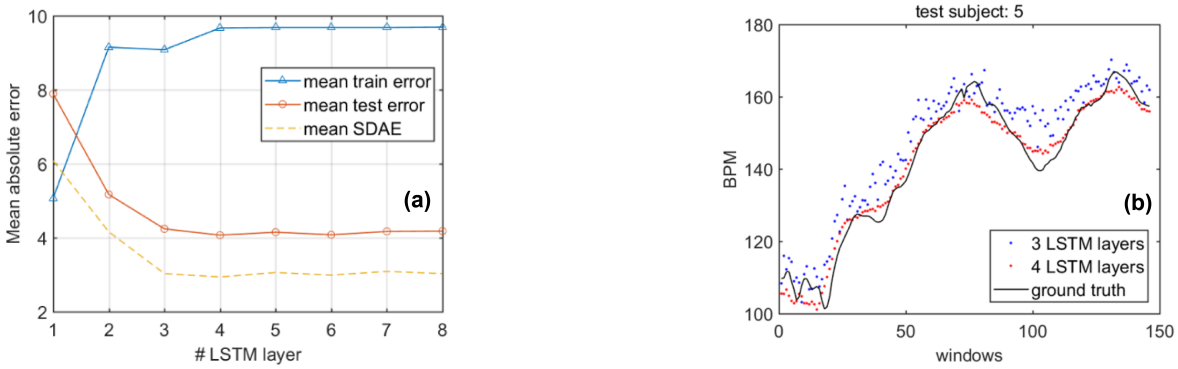

Fig. 3: Number of LSTM layers versus accuracy trade-off evaluation. (a) Impact of the number of layers on the model accuracy. (b) Influence of number of layers on predicted HR for IEEE SPC subject 5.

TABLE I: Accuracy (MAE±SDAE) results comparison w.r.t the literature.

PPG Subject	[1]CorNET [5]	Our LSTM-only network proposals Model I	Model II	Model III
1	3.76 ± 3.03	21.47±27.76	12.90±15.43	6.11±4.22
2	5.82 ± 10.96	17.63±21.47	10.65±12.65	2.87±2.32
3	1.47 ± 2.10	15.90±20.77	9.59±9.79	4.33±3.21
4	1.89 ± 5.77	13.62±20.03	4.77±6.15	3.84±3.21
5	1.04 ± 1.52	8.25±9.75	4.43±3.45	3.01±2.11
6	3.43 ± 11.85	14.04±20.19	8.07±10.57	3.83±3.40
7	1.08 ± 2.45	9.45±12.62	5.24±6.27	2.64±1.78
8	1.85 ± 6.96	17.21±17.98	12.88±16.27	11.42±10.20
9	1.74 ± 7.01	14.61±17.51	5.55±5.70	2.92±2.38
10	8.77 ± 10.10	15.32±18.58	5.56±6.59	6.88±6.44
11	3.39 ± 4.13	13.45±12.76	5.38±7.07	2.68±2.40
12	3.68 ± 6.31	17.67±15.63	5.25±6.17	3.14±2.55
Average	**3.16 ± 6.02**	**14.88±17.92**	**7.52±8.84**	**4.47±3.68**

[1]CorNET is the DNN architecture combining both CNN and LSTM.
[*]PP-Net is not considered as it is evaluated on a different dataset. As reference, the output is constrained to the 60-180 bpm interval.

MAE, but it is more suited for off-line processing as it needs to observe all input windows before any HR estimation. Real-time operation, required for edge devices, can only be achieved with models I and II, although the latter is preferred due to its considerably lower MAE.

Our LSTM-only network Model III proposal shows an accuracy penalty of 1.3 bpm compared to state-of-the-art works, although it has a much less complex model since it has 10× fewer parameters. Since the output is constrained to the 60-180 bpm interval, the Model III MAE has a relative error of up to 6.7% in the worst case.

The complexity advantage in terms of the complexity of LSTM-only networks can be verified in Table II. State-of-the-art deep learning-based implementations that employ a CNN+LSTM strategy need 4-10× more parameters per model and require up to 1000× more MAC operations per inferred output. Although the LSTM dataflow requires non-linear functions and feedback paths, it offers a viable alternative to other hybrid or purely feed-forward networks.

TABLE II: Complexity results comparison w.r.t the literature.

Model	Trainable Parameters	MACs	[†]Latency (samples)
CorNET [1]	256k	20.2M	1000
PP-Net [11]	124k	9.5M	250
Our Model I	7k	5.4M	1000
Our Model II	23k	17.4k	250
Our Model III	26k	3.7M	40k

[†]Minimum amount of processed input samples until first HR estimation.

Figure 4 shows the network performance for best and worst records given by subjects 11 and 9, respectively. In both cases, the ECG-based HR recording presented some extreme, short-period variations that the model could not capture, although most of these variations could be due to bad contact between the ECG patch and subject skin.

V. Conclusion

In this paper, an LSTM-only neural network for low-complexity HR estimation from wrist PPG is presented. Our exploratory approach evaluates the complexity-accuracy trade-off in this type of network. Compared to the state-of-the-art deep learning-based implementations, our proposal leads to a simpler model with 4-10× fewer parameters and up to 1000× fewer MACs per estimated output while attaining a similar prediction accuracy with an average error 8.7% lower than other implementations. Binary LSTMs and their effective RTL implementation have been reported in [5], [6]. This knowledge will be combined with our LSTM-only network to design a custom hardware accelerator.

References

[1] D. Biswas *et al.*, "CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 2, pp. 282–291, April 2019.

[2] T. Aoyagi and K. Miyasaka, "Pulse oximetry: Its invention, contribution to medicine, and future tasks," *Anesthesia and analgesia*, vol. 94, pp. S1–3, 02 2002.

[3] Z. Zhang *et al.*, "TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 522–531, Feb 2015.

[4] A. Temko, "Accurate Heart Rate Monitoring During Physical Exercises Using PPG," *IEEE Trans. on Biomedical Eng.*, vol. 64, no. 9, Sep. 2017.

[5] L. G. Rocha *et al.*, "Real-time HR Estimation from wrist PPG using Binary LSTMs," in *2019 IEEE Biomed. Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[6] ——, "Binary CorNET: Accelerator for HR Estimation From Wrist-PPG," *IEEE Trans. on Biomedical Circuits and Systems*, vol. 14, no. 4, pp. 715–726, 2020.

[7] D. Biswas, N. Simues-Capela, C. Van Hoof, and N. Van Helleputte, "Heart Rate Estimation From Wrist-Worn Photoplethysmography: A Review," *IEEE Sensors Journal*, pp. 1–1, 2019.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997.

[9] D. Velasco-Montero, J. Fernández-Berni, R. Carmona-Galán, and A. Rodríguez-Vázquez, "Optimum Selection of DNN Model and Framework for Edge Inference," *IEEE Access*, vol. 6, 2018.

[10] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi, and F. Kawsar, "Squeezing Deep Learning into Mobile and Embedded Devices," *IEEE Pervasive Computing*, vol. 16, no. 3, pp. 82–88, 2017.

[11] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "PP-Net: A Deep Learning Framework for PPG-Based Blood Pressure and Heart Rate Estimation," *IEEE Sensors Journal*, vol. 20, no. 17, 2020.

[12] S. Dieleman *et al.*, "Lasagne: First release," Aug. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.27878

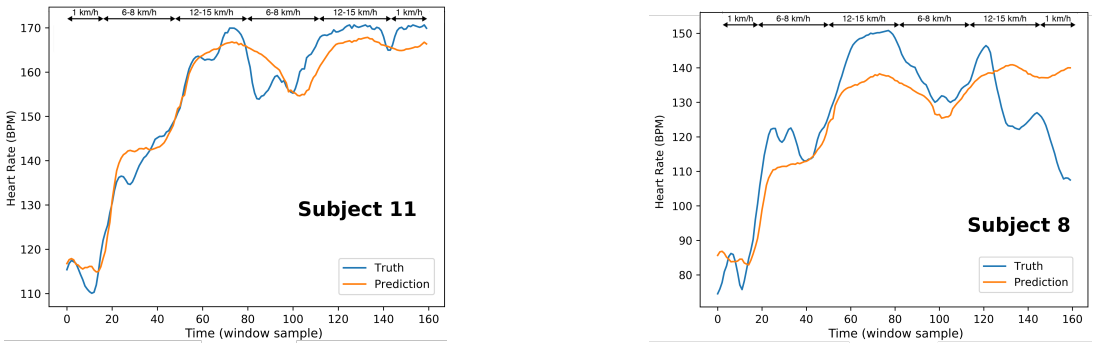

Fig. 4: PPG Estimated versus the true ECG HR for best (subject 11) and worst (subject 8) models.