2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Atrial Fibrillation Detection on Low-Power Wearables using Knowledge Distillation

Antonino Faraone[1], Halla Sigurthorsdottir[2], and Ricard Delgado-Gonzalo[2]

***Abstract*— The increasing complexity and memory requirements of neural networks have been slowing down the adoption of AI in low-power wearable devices, which impose important restrictions in computational power and memory footprint. These low-power systems are the key to obtain 24/7 monitoring systems necessary for the current personalized healthcare trend since they do not require constant charging. In this work, we apply Knowledge Distillation to our previously published convolutional-recurrent neural network for cardiac arrhythmia detection and classification. We show that the resulting network halves the memory footprint (138 K parameters) and the number of operations (1.84 MOp) compared to the baseline. By using Knowledge Distillation, this network also achieves significantly higher accuracy after quantization (increase in overall F_1 score from 0.779 to 0.828) and is capable of running into a nRF52832 System-on-Chip from Nordic Semiconductors. This promising result lays the groundwork for deployment on resource-constrained embedded platforms such as microcontrollers of the ARM Cortex-M family, thus potentially enabling continuous detection of cardiac arrhythmias in low-power wearable devices.**

I. Introduction

Deep Learning and Wearable Devices are extremely promising tools in the field of personal healthcare. On one side, with the emergence of big datasets, deep learning algorithms have achieved remarkable accuracy, progressively replacing traditional signal processing techniques in many application scenarios [1]. On the other side, low-power wearable devices have a big potential for personalized healthcare [2] as they allow for continuous detection of adverse medical conditions, disease, and emergency events [3], [4]. Note that the continuous non-interrupted monitoring plays a key role in finding important health-adverse events, thus leading to proper and fast diagnosis, as opposed to intermittent and on-demand monitoring mainstream devices such as the Apple Watch [5], [6]. The main issue arises due to the fact that deep neural networks are computationally complex and considerably resource-demanding (*e.g.*, memory, execution time). Therefore, in order to implement the inference in real time of neural networks on such resource-constrained devices, many compromises and optimizations are required in order to reduce the memory footprint and the computational complexity in order to achieve a real uninterrupted 24/7 monitoring during long periods of time.

[1]A. Faraone is with the Eidgenössische Technische Hochschule Zürich (ETHZ), Rämistrasse 101, Zürich, Switzerland (e-mail: afaraone@ethz.ch).

[2]H. Sigurthorsdottir and R. Delgado-Gonzalo are with the Centre Suisse d'Electronique et de Microtechnique (CSEM), Jacquet-Droz 1, Neuchâtel, Switzerland (e-mail: ricard.delgado@csem.ch).

This work focuses on the continuous monitoring of cardiac anomalies, more precisely, Atrial Fibrillation (AF) detection on low-power wearable devices. We focus on the the System-on-Chip (SoC) nRF52832 from Nordic Semiconductors, which is built around an ARM Cortex-M4f core, supports 512 kB of flash and 64 kB of RAM, since it offers a low-power consumption at a competitive price[1].

AF is the most common form of cardiac arrhythmia, and due to its paroxysmal and often asymptomatic nature [7], a device capable of continuously acquiring the user's electrocardiogram (ECG) and processing it to detect AF would be a great benefit for both doctors and patients. In the recent years, progress has been made in the field of AF detection on wearable devices, including some commercially available devices. However, high-power consumption of the algorithms employed is still a challenge for wearability and 24/7 continuous monitoring without recharging [8]. In our previous work, we introduced a wearable device capable of continuous AF monitoring from a single-lead ECG using a convolutional-recurrent neural network [9], which was in turn derived from a larger non-embeddable neural network [10]. Both networks were trained on a dataset provided for the Computing in Cardiology Challenge (CinC2017) [11]. The network in the wearable achieved, after quantization, an average F_1 score of 0.78 and a categorical accuracy of 0.85. Here, we extend upon this work to further decrease the computational complexity and the memory footprint of the network though a technique called Knowledge Distillation [12]. This process produces a new model that is faster to evaluate, and therefore deployable on less powerful hardware. Moreover, the student model tends to have greater generalization capabilities as well as a faster training. This technique has been successfully used in several applications of machine learning [13] such as object detection [14], acoustic models [15], and natural language processing [16].

The paper is organized as follows. In Section II, we describe the data that was used as well as the architecture and implementation of the distilled neural network into an embedded hardware. Then, in Section III, we evaluate its performance and robustness and compare the different models in terms of accuracy, memory footprint, and efficiency. Finally, in Section IV, we expose our main conclusions.

II. Materials and Methods

In this section, we first introduce describe the dataset that we used for training and validation. Then, we briefly

[1]https://infocenter.nordicsemi.com/pdf/nRF52832_PS_v1.4.pdf

show the architecture of the baseline network from [9] and present a smaller and less complex architecture (hereinafter *'slim network'*). We then describe the quantization technique applied to both models and finally the knowledge distillation technique used to retrain the slim network.

A. Dataset

We used the dataset from the challenge of CinC2017 [11]. It contains 8528 single-lead ECG signals recorded with an AliveCor device[2]. The signals are sampled at 300 Hz and have duration ranging from 9 to 60 seconds. All ECG signals are labeled with one of the following four classes: normal, sinus rhythm, atrial fibrillation, other rhythm, and noise. The proportion of each class in the dataset varies from 3.27% for noise to 59.52% for normal rhythm. In Figure 1, we show three ECG waveforms from the dataset in order to illustrate the signals that the network uses an input. Visually, we can observe that the normal beat (Figure 1a) is characterized for a regular rhythm of standard PQRST complexes, AF (Figure 1b) is characterized for irregular rhythm of standard PQRST complexes, and finally the noise class (Figure 1c) is characterized by random fluctuations of biopotential with possibility random R waves. It shall be noted that the labeling is not performed in a beat-to-beat manner. Instead, each label is rather assigned to the whole record. This generates some situations where a normal ECG may contain part of the recording corrupted by noise, but the ground truth assigns such recording to the noise class.

B. Network Architecture

1) Baseline Network: As a first approach, we employed our network described in [9], consisting in a convolutional block and a recurrent block. The former is a sequence of 7 1D-convolutional layers each followed by an average pooling layer with kernel size and stride equal to 2. Each convolution has a kernel size of 5, a padding of 2, and increasing number of filters. The recurrent block is a Gated Recurrent Unit (GRU) with 64 hidden units. The input signal is split into overlapping windows of 256 samples with a 50% overlap, then fed to the network one by one in sequence. Training is performed using Adam optimizer and we used the categorical cross-entropy as a loss function. This network network has 194 K parameters, and in order to process a window it requires 3.22 million operations (MOp), which translates into 0.09 s of execution time on the target platform (Nordic Semiconductor's nRF52832 SoC), using CMSIS-NN [17] as inference engine. As already noted in [9], most of the parameters and operations are accounted to the convolutional block.

2) Slim Architecture: Since the most resource-demanding part of the Baseline Network is the convolutional block, we reduced it in order to optimize its deployment and inference time into the selected SoC. This was achieved by reducing the depth of some filters and compensating the potential accuracy drop by adding a new layer. The resulting architecture

[2]https://www.alivecor.com/

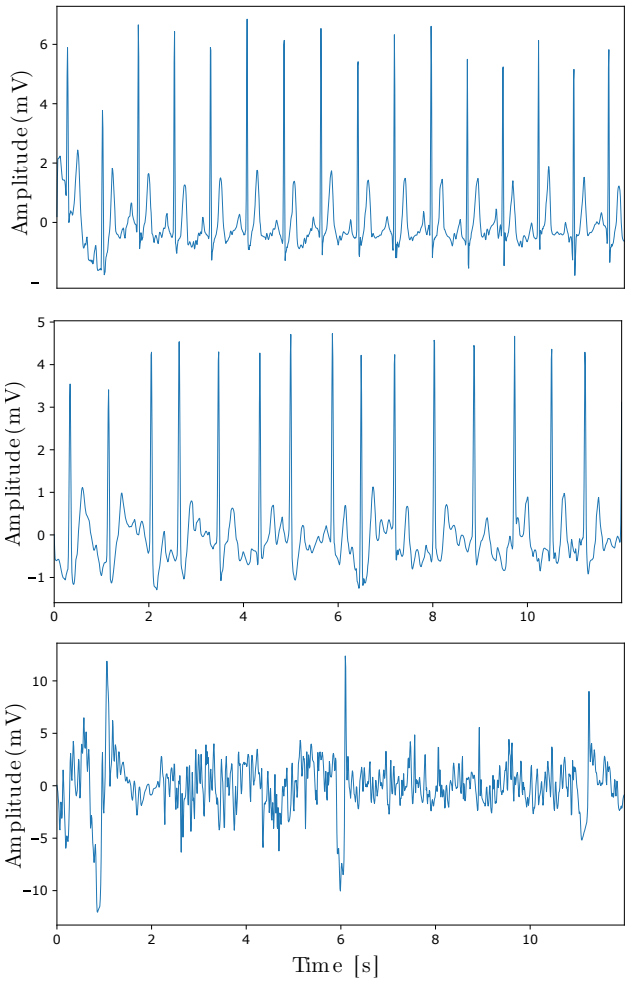

Fig. 1: Illustrative examples of some recordings from the dataset. (top) Normal beat. (middle) Atrial fibrillation. (bottom) Noise.

of the convolutional block is shown in Table I. Overall, the network has 138 K parameters (from which 100 K belong to the convolutional block), and requires roughly 1.84 MOp (0.06 s with the aforementioned hardware/software settings). As a consequence, it achieves an overall speed-up of 38% and a decrease in the memory footprint of 43% compared to the baseline.

C. Pre-Processing

The dataset was split into a training set of 7000 records and testing set of 1528 records. The data was pre-processed as described in [10]. To make it compatible with the sampling frequency of a novel wearable device for ECG recording under development in our research group, signals are resampled at 122 Hz. This value showed good empirical tradeoff between signal resolution and filter size, and consequently, memory footprint and computational load.

D. Q-format Quantization

Quantization is a technique that allows to map weights and activations, by default in floating-point format, into integers.

TABLE I: Architecture of the Slim Network and parameter count. N_w is the number of windows of the signal.

Layer	Output shape	Parameter count
Input	(N_w, 256, 1)	-
Conv1	(N_w, 128, 8)	48
Conv2	(N_w, 64, 16)	656
Conv3	(N_w, 32, 32)	2,592
Conv4	(N_w, 16, 32)	5,151
Conv5	(N_w, 8, 64)	10,304
Conv6	(N_w, 4, 64)	20,544
Conv7	(N_w, 2, 64)	20,544
Conv8	(N_w, 1, 128)	41,088
GRU	(64)	37,056
Dense	(4)	260
Total		138,244

The result is primarily a reduction in network's memory footprint, and a relaxation of computational effort, since operations are performed in integer-only arithmetic [18]. This approach is crucial for deployment to low power microcontrollers, since they have limited memory and often no native support to floating point operations. For the sake of compatibility with CMSIS-NN, we employed a symmetric fixed-point quantization where all the values are represented in $Qm.n$ format. In such quantization scheme, it follows that

$$q = \text{clip}(\text{round}(r \times 2^n))_{[-2^{m+n}, 2^{m+n}-1]}, \tag{1}$$

where q and r are the quantized and the real value respectively, and m and n are the number of bits allocated for the integer and the fractional part respectively. Experimentally, we found that $m = 2$ and $n = 5$ are suitable values to maintain accuracy.

E. Model Distillation

In order to improve the accuracy of the slim network, a technique called Model Distillation is employed. The idea, described by *Hinton et al.* in [12], is to use a large network (referred to as *the teacher*) to train a smaller network (*the student*). The purpose of this technique is to help the student network learn faster and with more generalization capabilities than direct training with on its own. To do so, the teacher is first trained using the target dataset as usual. Optionally, during this step, it is possible to set Softmax's temperature to a value larger than one. The Teacher is then used to generate the predictions for each element in the training set. Those predictions, called *soft labels*, are then used while training the student, instead of the ground truth (*hard labels*), to compute the loss.

In our work, our full model described in [10] is used as the teacher. The network contains more than 1 Million parameters, a larger input window (512 elements), and an LSTM with 128 hidden units per gate. The teacher is trained on the same dataset resampled at 200 Hz, using a temperature of 1 for the Softmax, since higher temperature values yielded poorer results. The student was the slim network described in Section II-B trained on the data as described in Section II-C and soft labels generated by the teacher.

III. Results

Multiple performance metrics of the discussed models are summarized in Table II. More precisely, we report the sensitivity and F_1 score metrics. The sensitivity, also known as recall, is defined as

$$s = \frac{tp}{tp + fn} \tag{2}$$

where *tp* is the number of true positives and *fn* is the number of false negatives. The F_1 score is the geometric mean of the sensitivity and precision, where the latter is defined as

$$p = \frac{tp}{tp + fp} \tag{3}$$

where *fp* is the number of false positives. In Table II, the best performance for each model is highlighted.

A. Baseline Network & Quantization

Focusing on the baseline, the achieved sensitivity of AF is 0.795, which remains stable after quantization, while the F_1 score, that was 0.775 in floating point precision, slightly decreases to 0.733. Robustness to Noise is the most affected by quantization, with a sensitivity drop of 16.6%. Overall, the network achieves an accuracy of 0.855 (almost unchanged by the reduced precision), and an average F_1 score of 0.792, that records a slight decrease to 0.772 after quantization.

B. Slim Network with Knowledge Distillation & Quantization

Due to the lower parameters count, as expected the slim network has slightly lower accuracy than the baseline network in detection of AF (-3% F_1 score). The other figures are not strongly affected by the reduced size of the network. After applying the technique described in Section II-E and performing quantization as described in Section II-D, we noticed an overall improvement in all the performance metrics, which are reported in the last column of Table II. In particular, distillation remarkably improved AF Sensitivity (+10.3%) and F_1 score (+8.6%). Compared to the baseline network, all figures improved. In particular, it shows a higher F_1 score of AF (+8.2%) and Other Rhythm (+5.6%), and higher sensitivity to noise (+30%), even after quantization. Overall, compared to the baseline the Distilled network has higher average Sensitivity (+7,8%) and F_1 score (+6.3%).

IV. Discussion and Conclusion

In this work, we presented an improved neural network for continuous Atrial Fibrillation detection that can run uninterruptedly on the low-power SoC nRF52832. The network is optimized to low-power platforms and it is more accurate than our network previously presented in [9]. The memory footprint of the new network is 43% smaller, thus compatible with devices with very limited memory space. Furthermore, after applying Knowledge Distillation and Quantization, the network performs better than the baseline in all metrics used (sensitivity and F_1 score), and far outperforms the baseline in AF detection and robustness to noise. The distillation process has allowed the slim network to obtain equivalent

TABLE II: Detailed performance metrics of the baseline before and after quantization. The best performance metric for each class is highlighted.

Class	Metric	Baseline	Baseline Quantized	Slim Network	Slim Network Distilled and Quantized
Normal Rhythm	Sensitivity	0.917	0.918	0.926	**0.934**
	F_1 score	0.915	0.916	0.920	**0.926**
Atrial Fibrillation	Sensitivity	0.795	0.810	0.748	**0.825**
	F_1 score	0.775	0.753	0.750	**0.815**
Other Rhythm	Sensitivity	0.761	0.752	0.770	**0.785**
	F_1 score	0.771	0.760	0.781	**0.803**
Noise	Sensitivity	0.705	0.588	**0.775**	0.765
	F_1 score	0.706	0.689	0.751	**0.770**
Overall	Sensitivity	0.794	0.767	0.805	**0.827**
	F_1 Score	0.792	0.779	0.800	**0.828**

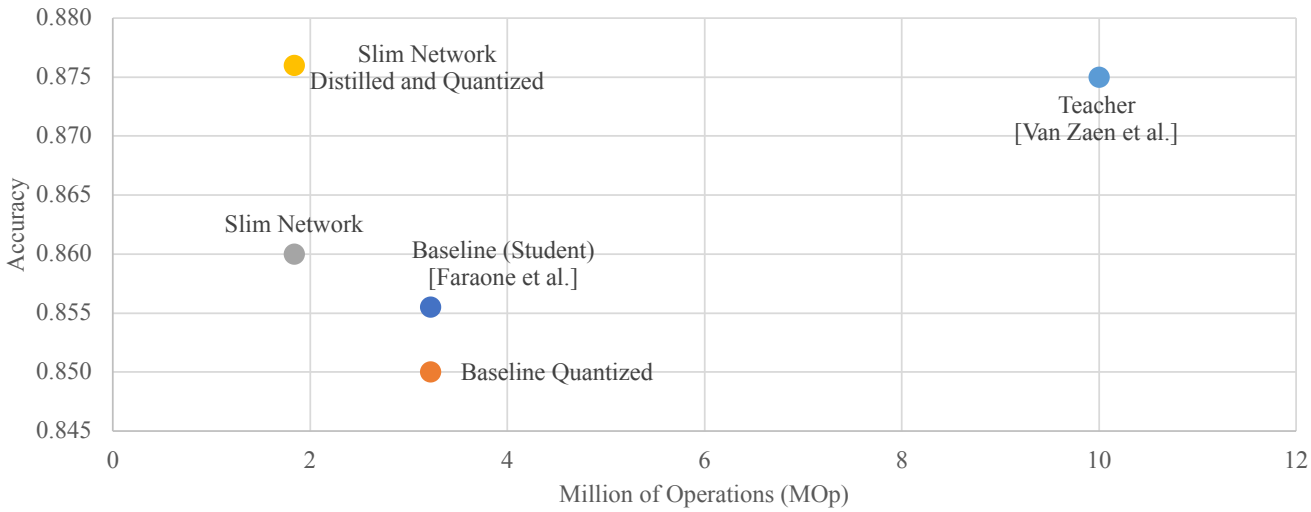

Fig. 2: Overall accuracy of the different neural networks versus the number of operations at inference time.

performance to the teacher at a fraction of computational resources and memory requirements. This showcased the fact that Knowledge Distillation is a suitable tool to reduce computational load in the process of deployment of NN to low power systems. Continuous monitoring on battery-powered wearables is an application scenario that, as showed in this work, could benefit from its adoption.

Future lanes of research include two major avenues. On one side, the presented network is deployable on existing wearable long-term monitoring systems such as the one described in [19]. This system was conceived for the European Space Agency for long-term monitoring of crew members. The embedding of the neural network would provide the system with diagnostic capabilities that were differed to offline processing due to memory and power requirements. Another extension avenue of this work is to increase the number of types of cardiac anomalies that the model could predict. For that purpose, we would base our developments on the new open dataset of the CinC2020 Challenge[3]. It includes multi-lead ECG data from the China Physiological Signal Challenge 2018[4] as well as a diverse population in the USA.

References

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, May 2017.

[2] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Personalized Medicine*, vol. 15, no. 5, pp. 429–448, Sep. 2018.

[3] C. Wang, W. Lu, M. R. Narayanan, S. J. Redmond, and N. H. Lovell, "Low-power technologies for wearable telecare and telehealth systems: A review," *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 1–9, Apr. 2015.

[4] J. Dieffenderfer, H. Goodell, S. Mills, M. McKnight, S. Yao, F. Lin, E. Beppler, B. Bent, B. Lee, V. Misra, Y. Zhu, O. Oralkan, J. Strohmaier, J. Muth, D. Peden, and A. Bozkurt, "Low-power wearable systems for continuous monitoring of environment and health for chronic respiratory disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1251–1264, May 2016.

[5] A. Khushhal, S. Nichols, Q. Evans, D. O. Gleadall-Siddall, R. Page, A. F. O'Doherty, S. Carroll, L. Ingle, and G. Abt, "Validity and reliability of the Apple Watch for measuring heart rate during exercise," *Sports medicine international open*, vol. 1, no. 06, pp. E206–E211, Oct. 2017.

[3] https://physionetchallenges.github.io/2020/

[4] http://2019.icbeb.org/Challenge.html

[6] N. Isakadze and S. S. Martin, "How useful is the smartwatch ECG?" *Trends in Cardiovascular Medicine*, vol. 30, no. 7, pp. 442–448, Oct. 2020.

[7] T. Yamashita, Y. Murakawa, K. Sezaki, M. Inoue, N. Hayami, Y. Shuzui, and M. Omata, "Circadian variation of paroxysmal atrial fibrillation," *Circulation*, vol. 96, no. 5, pp. 1537–1541, Sep. 1997.

[8] T. Pereira, N. Tran, K. Gadhoumi, M. M. Pelter, D. H. Do, R. J. Lee, R. Colorado, K. Meisel, and X. Hu, "Photoplethysmography based atrial fibrillation detection: A review," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–12, Jan. 2020.

[9] A. Faraone and R. Delgado-Gonzalo, "Convolutional-recurrent neural networks on low-power wearable platforms for cardiac arrhythmia detection," in *Proceedings of the 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS'20)*, Genova, Italy, Aug. 2020, pp. 153–157.

[10] J. Van Zaen, O. Chételat, M. Lemay, E. Muntané Calvo, and R. Delgado-Gonzalo, "Classification of cardiac arrhythmias from single lead ECG with a convolutional recurrent neural network," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSIGNALS'19)*, vol. 4, Prague, Czech Republic, Feb. 2019, pp. 33–41.

[11] G. D. Clifford, C. Liu, B. Moody, L.-W. H. Lehman, I. Silva, Q. Li, A. E. Johnson, and R. G. Mark, "AF classification from a short single lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017," in *Proceedings of the 2017 Computing in Cardiology (CinC'17)*, Rennes, France, Sep. 2017, pp. 1–4.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS'2015)*, 2015.

[13] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Honolulu, HI, USA, Jul. 2017, pp. 4133–4141.

[14] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'2017)*, Long Beach, CA, USA, Dec. 2017, pp. 742–751.

[15] T. Asami, R. Masumura, Y. Yamaguchi, H. Masataki, and Y. Aono, "Domain adaptation of DNN acoustic models using knowledge distillation," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, LA, USA, Mar. 2017, pp. 5185–5189.

[16] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, LA, USA, Mar. 2017, pp. 4825–4829.

[17] L. Lai, N. Suda, and V. Chandra, "CMSIS-NN: Efficient neural network kernels for ARM Cortex-M CPUs," *arXiv preprint arXiv:1801.06601*, 2018.

[18] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '18)*, Salt Lake City, UT,, Jun. 2018, pp. 2704–2713.

[19] O. Chételat, D. Ferrario, M. Proença, J.-A. Porchet, A. Falhi, O. Grossenbacher, R. Delgado-Gonzalo, N. Della Ricca, and C. Sartori, "Clinical validation of LTMS-S: A wearable system for vital signs monitoring," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'15)*, Milan, Italy, Nov. 2015, pp. 3125–3128.