Energy Efficient Software-hardware Co-design of Quantized Recurrent Convolutional Neural Network for Continuous Cardiac Monitoring

Jinhai Hu[1,2], Cong Sheng Leow[2], Wang Ling Goh[1] and Yuan Gao[2]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Republic of Singapore
[2]Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Singapore 138634, Republic of Singapore
Email: jinhai001@e.ntu.edu.sg, leow_cong_sheng@ime.a-star.edu.sg, ewlgoh@ntu.edu.sg, gaoy@ime.a-star.edu.sg

***Abstract*—This paper presents an electrocardiogram (ECG) signal classification model based on Recurrent Convolutional Neural Network (RCNN). With recurrent connections and data buffers, a single convolutional layer is reused to implement multiple layers function. Using a 5-layers CNN network as an example, this approach reduces the number of parameters by more than 50% while achieving the same feature extraction size. Furthermore, quantized RCNN (QRCNN) is proposed where the input signal, interlayer output, and kernel weights are quantized to unsigned INT8, INT4, and signed INT4 respectively. For hardware implementation, pipelining and data reuse within the 1-D convolution kernel can potentially reduce latency. QRCNN model achieved 98.08% validation accuracy on MIT-BIH datasets with only 1% degradation due to quantization. The estimated dynamic power consumption of the QRCNN is less than 60% of a conventional quantized CNN when implemented on a Xilinx Artix-7 FPGA, showing the potential for resource-constraint edge devices.**

Keywords—RCNN, Fixed-point quantization, FPGA, ECG classification

I. Introduction

Early detection and prediction based on real-time processing are critical to improving the quality of life for patients with cardiac diseases. Cardiovascular diseases (CVDs) such as heart failure and atrial fibrillation are the leading causes of death globally [1]. Electrocardiogram (ECG) is the gold standard for cardiac monitoring by measuring the strength and timing of the electrical activity in the heart [2]. However, some cardiovascular events such as heart rate variability (HRV) and atrial fibrillation are rare and not predictable, which cannot be detected by regular health checks [3]. Therefore, long-term ECG monitoring is necessary to capture such patterns for further evaluation. Currently, patients are required to wear a Holter monitor which sends recorded data to the cloud for offline analysis. Wearing the Holter monitor for long period may cause discomfort or skin irritation, and the sensors can be interrupted by other household appliances [4]. Thus, there is a rising need for instant ECG prediction and classification solutions for shorter diagnosis periods and greater comfort.

To classify a time-variant sequence such as an ECG signal, the classifier needs to identify the key features from the data series and map them to the respective pattern categories. There are a few key ECG anomaly patterns illustrated in Fig. 1, including Supraventricular premature beat (S), Premature ventricular beat (V), Fusion of the ventricular and normal beat (F), and Unclassifiable beat (Q). Together with the Normal beat (N), five categories suggested by the AAMI standard [5] are required for meaningful ECG classification.

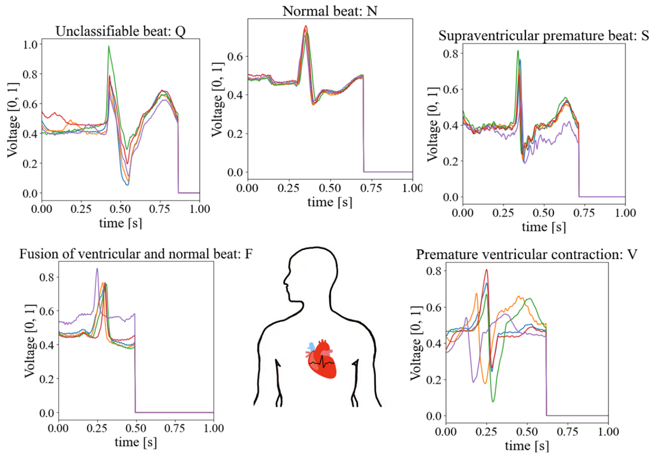

Fig. 1. Illustration of ECG signal with five key types of waveform pattern (normalized).

Many prior works focused on ECG signal processing, through methods ranging from conventional analysis to deep learning algorithms. For instance, [6] proposed a processor to detect shockable cardiac arrhythmias with a sensitivity of 98%. On the other hand, deep learning approaches, such as recurrent neural networks (RNNs), have become increasingly popular for ECG forecasting and classification. Long Short-Term Memory (LSTM) [7] and Gated Recurrent Unit (GRU) [8] have been shown to perform well for time-series data such as ECG. Although RNNs can give real-time outputs, deeper networks such as bidirectional LSTM are required to reach high accuracy on practical ECG applications, making them unsuitable for edge devices due to their high compute requirements. A dynamically-biased LSTM is proposed in [9] to improve classification accuracy with shorter window sizes and fewer training parameters. Inspired by the visual-cortical system, Convolutional Neural Networks (CNNs), both one-dimensional (1D) [10-13] and two-dimensional (2D) [14-16] have also shown good potential in ECG analysis. In 2D-CNN, the ECG signal is converted to images and then perform image processing. To obtain satisfactory results, large-scale deep 1D- and 2D-CNN are required. For example, the 2D-CNN model in [14] contains more than 10 million weight parameters and requires huge computation power. While 1D-CNN [11] includes multiple layers and results in high computation latency. Therefore, it is highly challenging to implement deep

This work was supported by the Agency for Science, Technology and Research (A*STAR), Singapore under the Cyber-Physiochemical Interface programme, grant No. A18A1b0045.

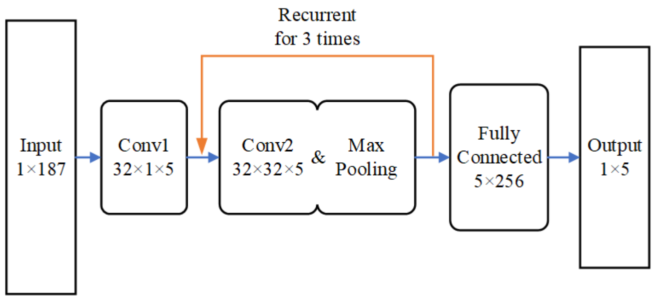

Fig. 2. Block diagram for Recurrent Convolutional Neural Network (RCNN).

learning approaches such as RNN and CNN in resource-constraint devices.

In this paper, an improved Recurrent CNN (RCNN) model is proposed for energy-efficient ECG classification through convolution layer reusing. Through recurrent connections, the output from the first convolution layer is passed through the same second layer four times, which reduces the number of training parameters significantly. The overall network is effectively reusing 2 convolution layers to implement operations equivalent to 5 convolution layers. The Quantized RCNN (QRCNN) model with reduced precision is suitable for hardware implementation. The QRCNN obtained 98.08% validation accuracy for 5-category classification with the MIT-BIH ECG dataset [17] at less than 6600 parameters. Pipelining within the convolution kernel supports QRCNN for real-time processing and reduces latency in FPGA inference. This paper is organized as follows: Section II introduces the proposed model, fixed-point model truncation, and hardware simulation. The results of ECG classification and benchmarking are discussed in Section III, followed by a summary in Section IV.

II. Proposed Model

A. *Model Architecture*

The proposed RCNN ECG classification algorithm block diagram is illustrated in Fig. 2. The model consists of five layers, namely the input layer, two 1D CNN layers (Conv1 and Conv2), one Fully Connected (FC) layer, and an output layer. The input data only need to go through a single step of pre-processing, that is min-max normalization shown in Fig. 1. Compared to z-score normalization, min-max normalization is easy to realize at the front-end level of the ECG signal recording system. The ECG signal is isolated into heartbeat segments based on Christov's peak detection algorithm [18]. With time-series ECG, the 1D CNN is effective for examining features within shorter segments of the overall dataset, where the location of features within the segments is not of high relevance [19]. As the size of CNN is highly related to the input sequence length, the length of one full heartbeat is reduced to half before sending to the Conv1 layer. Conv1 layer extracts the single-feature ECG data into 32 channels with a kernel size of 5 and without padding. The 32 features are then sent to a recurrent block consisting of the Conv2 layer with max-pooling for 4 times, which outputs the same number of channels to feedback into the same Conv2 layer (same weight). As our preliminary exploration indicated the optimal network with 5 convolution layers, the recurrent count is set to 4 in the proposed solution to achieve equivalent result. The max-pooling layer is the key component to reduce the size of features, and padding is added to Conv2 to prevent features from losing margin information.

The recurrent block provides several advantages. First, it maintains high accuracy while simplifying the model. The performance will degrade significantly if a larger stride and

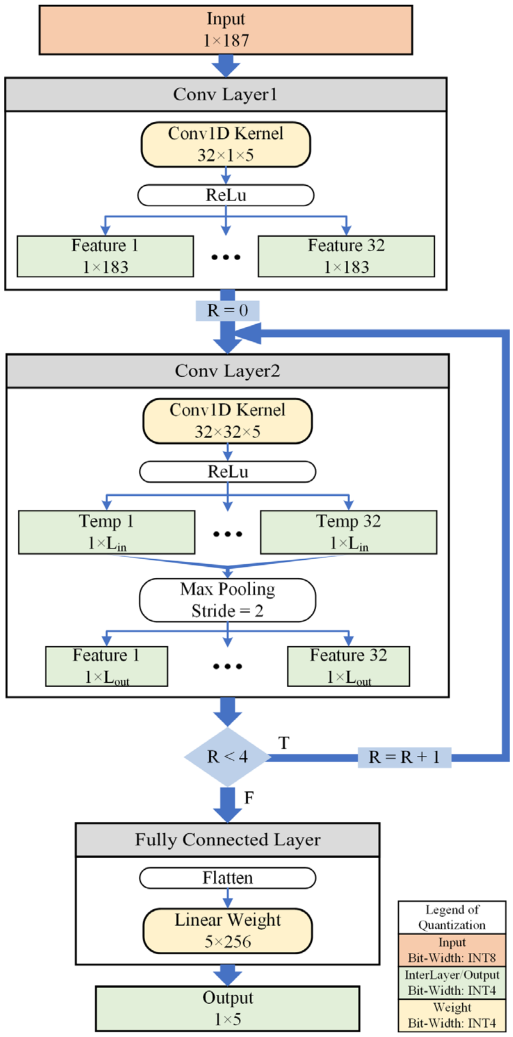

Fig. 3. Detailed diagram for QRCNN model, where orange, green and yellow blocks stand for unsigned INT8, unsigned INT4 and signed INT4 correspondingly.

kernel of the max-pooling layer is set since too much information is lost without feature extraction via the recurrent Conv2 layer. Furthermore, the batch normalization layer can be omitted since the training process of the recurrent block will reduce the effect of internal covariate shift caused by multiple different hidden layers. That is, each input will pass through an equivalent of 6 layers, of which only 3 (Conv1, Conv2, and FC) requires weight updates during training. At the same time, it saves 70% of trainable parameters by reusing the same Conv2 kernel four times, which will consume less memory when implement on hardware.

After the recurrent blocks, the 32 short but essential features are flattened and then go through a fully connected layer for classification. The optimized RCNN for ECG classification obtained total weight parameters of 5280 and 1280 for CNN layers and fully connected layers, respectively.

B. *Fixed-Point Quantization*

The weights in the AI algorithms are normally represented in FP32. However, as the goal of this work is to implement the algorithm in portable devices, quantization is required for all weight parameters, inputs, and interlayer outputs. The detailed QRCNN model is shown in Fig. 3, where three types of quantization are applied to different data variables using quantization-aware training (QAT). The input ECG signal is truncated into unsigned INT8 since it has already been normalized into 0 to 1. All the weights are binarized and

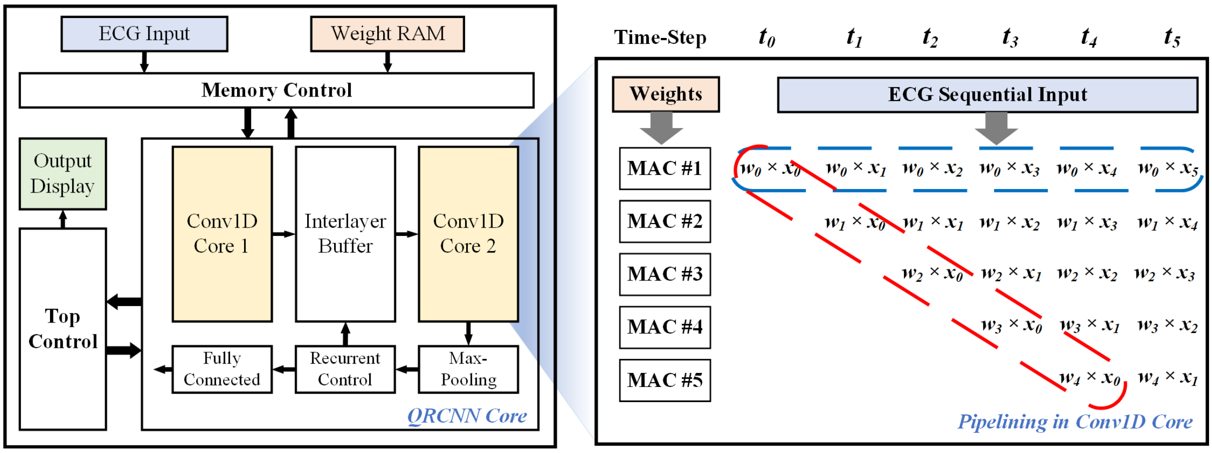

Fig. 4. Schematic diagram for QRCNN with two 1D-convolutional cores involving pipelining. (left) Data flow in single Conv1D core, where red and blue dash blocks stand for the reusing of input data and kernel weight. (right)

TABLE I MULTI MODEL PERFORMANCE COMPARISON

Topology	Number of CNN Layers	Number of Recurrent Loops	Number of Weight Parameters	Validation Accuracy	Weight Quantization
CNN	5	0	21,920	99.40%	FP32
QCNN	5	0	21,920	98.52%	INT4
RCNN	2	4	6,560	99.10%	FP32
QRCNN	2	4	6,560	98.08%	INT4

quantized into signed INT4 with the same scaling of 0.0625 (2^{-4}), where all trainable parameters can only be selected from 15 choices from -0.375 to +0.375. Therefore, the restriction of weight quantization can further reduce the problem of internal covariate shift since gradient vanishing is limited by the lower and upper bond of weights, which is an additional justification for eliminating batch normalization. Moreover, biases are removed from CNN kernels and fully connected layers since the effect of adjusting the activation function did not contribute to classification accuracy significantly [20]. Reducing bias can reduce the number of adders, benefitting the convolution pipeline design, without affecting the scaling of interlayer output. Since ReLU is followed by each Conv layer, all interlayer outputs will be non-negative numbers. As a result, the quantization for interlayer outputs is set as unsigned INT4.

Inside the Conv Layer2 block in Fig. 3, the output sequence length (L_{out}) is expressed as (1)

$$L_{out} = \left\lfloor \frac{L_{in} - L_{kernel}}{S_{pooling}} \right\rfloor + 1 \quad (1)$$

where L_{in} is the input sequence length in Conv Layer2, K_{ernel} is the max-pooling kernel size which is set as same as Conv kernel size and $S_{pooling}$ is the stride of the max-pooling layer.

Thus, QRCNN is a fully quantized model trained through QAT, which minimizes the degradation on edge devices. All fine-trained weights are exported after software training and are to be imported to FPGA memory.

C. Hardware Implementation

Based on the QRCNN architecture discussed in the previous section, we estimate the performance of the recurrent architecture on a Xilinx Artix-7 FPGA, whose overall schematic diagram is shown in Fig. 4. Some of the metrics of interest include the resource utilization and the power consumption of the blocks, which are paramount in realizing cardiac monitoring at the edge. The base unit for the

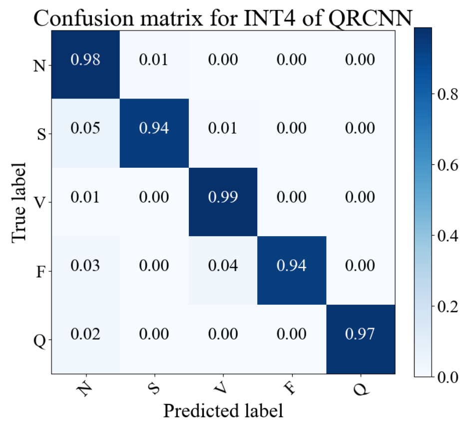

Fig. 5. Block diagram for Recurrent Convolutional Neural Network.

convolution block is the 1D-Convolutional Core (Conv1D Core) which calculates the output of the convolution between the kernel and the input data. To reduce latency and resources, a pipelining architecture is implemented for the Conv1D Core based on the paradigms proposed by [21]. Fig. 4 (right) shows the pipelining system used in Conv1D Core, where each data point is fed in only once due to data reuse. Using the same software architecture, the hardware architecture uses only 2 Conv1D Cores.

Additionally, the kernel size in QRCNN determines the multiply-and-accumulate (MAC) units as seen in Fig. 4, each tied to the corresponding weight. During each clock cycle, data is fed into all the MAC units, which will multiply the data and the weight before summation. Once the convolution data is ready for that cycle, the READY status will be triggered; the FINISH status will be triggered when the full convolution cycle consisting of n_{out} cycles has been completed, where n_{out} is determined via (2). The in-built buffer within each Conv1D Core allows the finished output to be consolidated and serialized for the next core, while the FINISH signal serves as the RESET signal for the next core.

$$n_{out} = \left\lfloor \frac{n_{in} + 2p - k}{s} \right\rfloor + 1 \quad (2)$$

where n_{in} is the input size, p is the padding size, k is the kernel size, and s is the stride length.

III. RESULTS AND DISCUSSION

A. Simulation and Hardware Results

The performance of the proposed algorithm is evaluated with the MIT-BIH ECG database which includes 48 sets of

TABLE III BENCHMARK TABLE

	TBioCAS 2019 [11]	OJCAS 2021 [22]	ISCAS 2021 [23]	TBioCAS 2022 [24]	JBHI 2022 [20]	ISCAS 2023 [25]	TBioCAS 2022 [26]	**This Work**	
Dataset	MIT-BIH	MIT-BIH	PTB Diagnostic	MIT-BIH	MIT-BIH	MIT-BIH	CHB-MIT	**MIT-BIH**	
Topology	CNN+MLP	AE-ELM	LSTM	Binary CNN	CNN+MLP	LSTM	CNN	**RCNN**	
Features	RR Interval	RR Interval	RR Interval	Binary Images	RR Interval	RR Interval	Compressed Signal	**RR Interval**	
No. of Weight Parameters	198,037	72,400	NA	65,729	8,157	8,608	≈16k*	**6,560**	
No. of Classes	5	2	2	2	5	5	2	**5**	
Validation Accuracy	98.4% (Software)	92.0%	84.6%	97.3%	99.1%	96.7%	89.7%	**98.08%**	**99.10%**
Weight Resolution	INT8	INT8	INT1	INT2	INT8	INT4	INT3	**INT4**	**FP32**
Activ-Function Resolution	INT8	INT8	INT5	INT2	INT8	FP32	INT8		
Interlayer Output Resolution	INT8	INT8	INT5	INT2	INT8	FP32	INT8		
Hardware Type	MCU	ASIC	FPGA	FPGA	MCU	FPGA	FPGA	**FPGA**	
Clock Frequency	NA	70MHZ	NA	100KHz	NA	40KHz	100MHz	**100KHz**	
LUTs	NA	NA	NA	2873	NA	3980	35631	**1060**	
Registers	NA	NA	NA	2104	NA	2376	8672	**1825**	
Dynamic Power (μW)	141.4	2940	NA	404	Normal: 400 Abnormal: 3k	40	50	**15**	

* Calculated based on information provided in paper.

TABLE II MULTI HARDWARE TOPOLOGY COMPARISON

Topology	LUTs	Registers	Dynamic Power (μW)	Total On-chip Power (mW)
QCNN	1655	1854	25	91.025
QRCNN	**1060**	**1825**	**15**	**91.015**

30-minute-long ECG data. Each dataset contains both normal and abnormal ECG features (N, S, V, F, Q) recorded with a fixed sampling frequency of 360 Hz and 11-bit data resolution.

After pre-processing, the training, testing, and validation dataset contains 80%, 10%, and 10% of the whole data, respectively. Based on the same validation set, Table I shows the model comparison between conventional CNN and RCNN including quantized performance, while Fig. 5 shows the confusion matrix of the QRCNN classification result. The proposed RCNN model achieved similar accuracy as conventional CNN but consumed less than 30% of CNN training parameters. However, since the degradation due to quantization is more sensitive to smaller weight sizes [27, 28], the loss of accuracy of RCNN after model truncation is slightly higher than the conventional CNN using (loss function).

Table II shows the estimated resource utilization for the implementation of a different number of Conv1D Cores on Xilinx's Artix-7 at 100KHz with 4-bit precision. When using the Conv1D Cores for QRCNN, the increase in resources required for control is significantly less than the increase brought by using more layers. Based on the comparison of the proposed QRCNN and the conventional QCNN using 5 layers, a 36% reduction in LUT is possible, with a 40% reduction in power. While some basic pipelining has been implemented to reduce resources and power, further optimization can be done to further reduce the utilization and power. One potential approach is to implement power gating such that the sparsity of the kernel can be fully utilized for reduced power consumption. More efficient MAC units and optimized pipelining can potentially reduce the overall resource and power footprint. Nonetheless, the significant reduction in both resource and power when using the proposed QRCNN as compared to the conventional QCNN approach demonstrates the viability of QRCNN as the better architecture for hardware implementation at the edge.

B. Benchmarking and Discussion

The performance of the proposed model is summarized and compared with other state-of-arts in Table III. Both software and hardware evaluation aspects are included in the benchmark table. Compared to those designs implemented at FP32 resolution, the proposed algorithm used one of the smallest network sizes and achieved comparable accuracy. This model is satisfactory with minimal pre-processing, which proves the robustness of RCNN. [25] achieved 96.7% accuracy under around 8600 trainable weight parameters with INT4 quantization. However, sigmoid and tanh activation functions in [25] are necessary logics in LSTM, whose quantization will degrade its nonlinearity significantly. When the model is fully quantized to INT4 instead of truncating weight resolution alone, the accuracy only has minor degradation of 1.02%. Together with layer reuse, the pipelining dataflow causes higher dynamic power but reduces the processing latency.

IV. CONCLUSION

In this paper, an ECG classification model based on RCNN is proposed for resource-constraint edge devices. RCNN can effectively extract the classification information with reduced overall network computation complexity. The proposed method achieved 99.1% accuracy for five categories of ECG classification with the MIT-BIH database. The model is further improved to a fully quantized model, QRCNN, and reached 98.08% validation accuracy. Moreover, this model consumed fewer hardware resources when implemented on FPGA. The next step is to implement this real-time classification algorithm in CMOS ASIC format to provide ECG anomaly classification for continuous cardiac monitoring. Furthermore, further analysis of latency could be conducted on different devices to observe the effects of this algorithm on the latency.

REFERENCES

[1] S. Mendis, P. Puska, and B. Norrving, Global atlas on cardiovascular disease prevention and control. *World Health Organization*, 2011.

[2] L. S. Lilly, Pathophysiology of heart disease: A collaborative project of medical students and faculty, 6th ed. *Wolters Kluwer*, 2016.

[3] Virani SS, *et al*, "Heart disease and stroke statistics-2021 update: a report from the American Heart Association" *Circulation*. vol. 143, no.8, pp. 254-743, 2021.

[4] D. Rojers, *et al*, Holter Monitor. *American Heart Association*, 2015.

[5] ANSI/AAMI "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," *Assoc. Adv. Med. Instrument*, no. EC57, 1998.

[6] S. M. Abubakar, *et al*, "A wearable auto-patient adaptive ECG processor for shockable cardiac arrhythmia," *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Taiwan, 2018, pp. 267-268.

[7] Hochreiter, S. and J. Schmidhuber. "Long short-term memory," *Neural Comput. Appl.*, vol. 9, no. 8, pp.1735-1780, 1997.

[8] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS Workshop on Deep Learning*, December 2014.

[9] J. Hu, W. L. Goh and Y. Gao, "Dynamically-biased fixed-point LSTM for time series processing in AIoT edge device," *IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, pp. 1-4, 2021.

[10] U. R. Acharya, *et al*., "A deep convolutional neural network model to classify heartbeats," *Comput. Biol. Med.*, vol. 89, pp. 389-396, 2017.

[11] N. Wang, *et al*, "Energy-efficient intelligent ECG monitoring for wearable devices," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 1112-1121, Oct. 2019.

[12] X. Li, *et al*, "A 1D convolutional neural network for heartbeat classification from single lead ECG," *IEEE Int. Conf. Electron., Circuit. Syst. (ICECS)*, pp. 1-2, 2020.

[13] M. Kachuee, S. Fazeli and M. Sarrafzadeh, "ECG heartbeat classification: a deep transferable representation," *IEEE Int. Conf. on Healthcare Informatics (ICHI)*, New York, NY, USA, 2018, pp. 443-444.

[14] C. Cordoș, *et al*, "ECG signal classification using convolutional neural networks for biometric identification," *IEEE Int. Conf. Telecom. and Signal Process. (TSP)*, pp. 167-170, 2021.

[15] C. Szegedy, *et al,* "Rethinking the inception architecture for computer vision", *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2818-2826, 2016.

[16] S. M. Abubakar, *et al*, "A 746 nW ECG processor ASIC based on ternary neural network," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 4, pp. 703-713, Aug. 2022.

[17] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. in Med. and Biol.*, vol. 20, no. 3, pp. 45-50, May-June 2001.

[18] C. Ivaylo, "Real time electrocardiogram QRS detection using combined adaptive threshold," *Biomedical Engineering*, 2004.

[19] S. Kiranyaz, T. Ince and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. on Biomed. Eng.*, vol. 63, no. 3, pp. 664-675, March 2016.

[20] J. Xiao, *et al*., "ULECGNet: An ultra-lightweight end-to-end ECG classification neural network," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 206-217, Jan. 2022.

[21] Y. -H. Chen, *et al*, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[22] Y. -C. Chuang, *et al*, "An arbitrarily reconfigurable extreme learning machine inference engine for robust ECG anomaly detection," *IEEE Open Journal of Circuits and Systems*, vol. 2, pp. 196-209, 2021.

[23] S. A. Mirsalari, *et al*, "ELC-ECG: Efficient LSTM cell for ECG classification based on quantized architecture," *IEEE Int. Symp. Circuits Syst (ISCAS)*, pp. 1-5, 2021.

[24] D. L. T. Wong, *et al*, "An energy efficient ECG ventricular ectopic beat classifier using binarized CNN for edge AI devices," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 2, pp. 222-232, April 2022.

[25] J. Hu, *et al*, "Classification of ECG anomaly with dynamically-biased LSTM for continuous cardiac monitoring," *IEEE Int. Symp. Circuits Syst. (ISCAS)*, accepted, 2023.

[26] D. Wu, S. Zhao, J. Yang and M. Sawan, "Software-hardware co-design for energy-efficient continuous health monitoring via task-aware compression," *IEEE Trans. Biomed. Circuits Syst. (TBioCAS)*, early access, 2023.

[27] W. Nogami, *et al*, "Optimizing weight value quantization for CNN inference," *International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-8.

[28] S. I. Young, *et al*, "Transform quantization for CNN compression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5700-5714, Sept. 2022.