An Energy-Efficient Configurable Coprocessor Based on 1-D CNN for ECG Anomaly Detection

Chen Zhang, Zhijie Huang, QianXi Cheng, Changchun Zhou and Xin’an Wang
School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Shenzhen 518055, China
Email: {zhangchenn, zhouchch, anxinwang}@pku.edu.cn, {zhijiehuang, qxcheng}@stu.pku.edu.cn

***Abstract*—Many healthcare devices have been widely used for electrocardiogram (ECG) monitoring. However, most of them have relatively low energy efficiency and lack flexibility. A novel ECG coprocessor is proposed in this paper, which can perform efficient ECG anomaly detection. In order to achieve high sensitivity and positive precision of R-peak detection, an algorithm based on Hilbert transform and adaptive threshold comparison is proposed. Also, a flexible one-dimensional convolutional neural network (1-D CNN) based classification engine is adopted, which can be configured with instructions to process various network models for different applications. Good energy efficiency is achieved by combining filter level parallelism and output channel parallelism within the processing element (PE) array with data reuse strategy. A 1-D CNN for arrhythmia detection is proposed to validate the hardware performance. The proposed ECG coprocessor is implemented using 55 nm CMOS technology, occupying an area of 1.39 mm^2. At a clock frequency of 100MHz, the energy efficiency is 215.6 nJ/classification. The comparison results show that this design has advantages in energy overhead and detection performance.**

Keywords—ECG, coprocessor, classification, R-peak, 1-D CNN

I. Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide, seriously affecting the life quality of patients and putting pressure on the social healthcare system. ECG monitoring is an essential method for early prevention of CVD. Currently, portable ECG monitoring devices have been widely used in households. However, these devices typically rely on remote analysis and require binding to mobile phones and dedicated apps, which is inconvenient for elderly users. Remote analysis requires high wireless network quality, which limits the availability of devices in certain scenarios. Also, for users with a clear risk of CVD, remote ECG analysis may not be timely enough in emergency situations. To enable local analysis on edge devices, many recent studies have proposed ECG processors.

Neural network (NN) based classifiers have been widely used in ECG processors. In [1], the ECG is digitized by a level crossing ADC, and the converted data is fed to an artificial neural network (ANN) for classification, achieving an accuracy of 98%. A deep neural network (DNN) based arrhythmia classifier is proposed in [2], the DNN has two hidden layers and can classify ECG into 5-class with an accuracy of 91.6%. A biomedical AI processor with a NN engine is proposed in [3], which utilizes an adaptive learning architecture to address changes between patients, event-driven method is used to reduce power. A sparse CNN accelerator for ECG classification is proposed in [4], four 12-stage cascade PE structures are used to process the CNN. An inference engine based on extreme learning machine (ELM) is presented in [5], which can classify ECG into two types with an accuracy of 92%. However, due to the lack of configurability and relatively low energy efficiency, the application of previous work is limited. On the one hand, local analysis requires flexibility to adapt the individual variations in ECG signals from users and different abnormal ECG detection. On the other hand, to increase the battery life and achieve long-term healthcare, energy consumption is required to be low. In addition, considering the classification of multi-lead ECG signals, high throughput is also necessary.

To overcome existing problems, a flexible ECG coprocessor for healthcare devices is proposed in this paper. The main contributions are as follows: 1) We propose a R-peak detection method with high sensitivity (Sen) and positive precision (+P). Data-driven dataflow is adopted to implement the detection engine; 2) A lightweight 1-D CNN ECG classification model is proposed, which can classify 5 types of arrhythmias with an accuracy of 97.7%; 3) An architecture for 1-D CNN acceleration is proposed, which achieves high energy efficiency through data reuse and time multiplexing strategies; 4) We design a configurable central controller that provides flexible computing support for different 1-D CNNs for ECG anomaly detection through external write instructions. This paper is organized as follows: Section II depicts the architecture and the details of the modules. Section III presents the hardware implementation and results. Finally, the conclusions are given in Section IV.

II. System Architecture

A. The Overall Architecture

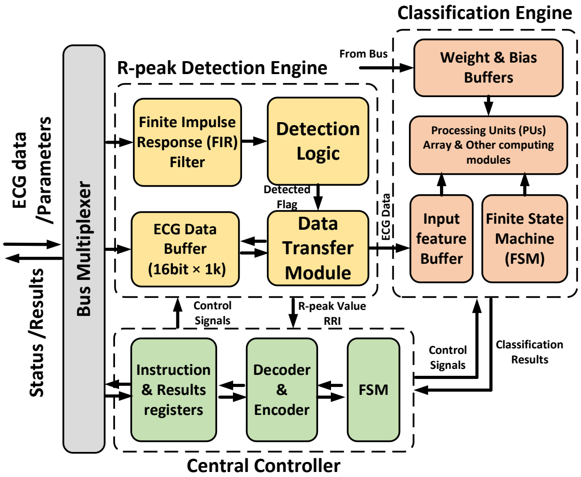


Fig. 1. The proposed ECG coprocessor architecture.

The architecture of the proposed 1-D CNN based ECG coprocessor is shown in Fig. 1. It consists of an R-peak

This work was supported by Shenzhen Science and Technology Innovation Commission (Project No. KQTD20200820113105004).

detection engine, data transfer module, classification engine, and central controller. Digitized ECG data is input to the coprocessor through a customized interface, and parameters of CNN and instructions are also transmitted through this interface. Once an R-peak is detected, the index address of peak value in ECG data buffer will be calculated and saved in a register that stores 3 R-peak indexes. When saving the latest index, the oldest index will be discarded. To ensure that ECG data containing complete R-peak can be obtained for subsequent classification, the data points centered around the second R-peak index will be sent to the input feature buffer by the data transfer module. In addition, the RR interval (RRI) is also calculated using the index values, and for a fixed sampling rate, heart rate (HR) can be easily obtained through RRI.

B. *R-Peak Detection Engine*

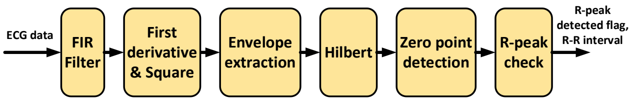


Fig. 2. The proposed R-peak detection logic.

The logic diagram of the R-peak detection engine is shown in Fig. 2. Firstly, a 69-tap Finite Impulse Response (FIR) filter is used to remove the noise in original digitized ECG signal such as power line interference, baseline drift, electromyography (EMG) noise, and motion artifacts. Considering the frequency range of the QRS complex, the bandwidth is set to 2-28 Hz. The filtering operation is data driven, the earliest data will be discarded when new data arrives, and the convolution is enabled simultaneously，as shown in Fig. 3. Only one multiplier accumulator (MAC) is used to complete all filtering convolution, which greatly reduces hardware overhead.

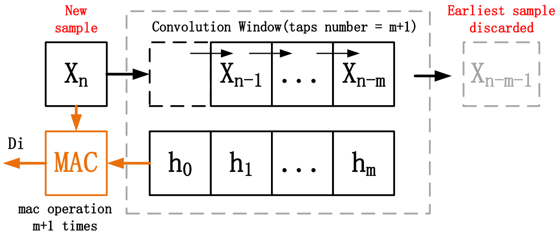


Fig. 3. The data driven filtering convolution.

Then first-order derivative and squared calculations are performed on the filtered data to highlight the R-peak feature and suppress noise signals, as follows:

$$S(i) = \left(D(i) - D(i-1)\right)^2 \tag{1}$$

Triangular window filtering is used for envelope extraction to smooth signals and remove noise. It is a linear filtering method based on sliding windows. The triangular window is defined as:

$$w(n) = (1 - |(n-2N)/2N|) \times a, 1 \le n \le 2N-1 \tag{2}$$

where w(n) is the value of the window function at index n, 2N-1 is the window length and a is the scaling factor. The signal sequence x(n) is convolved with the triangular window function to obtain the filtered output signal y(n). Convolutional operations can be expressed as:

$$y(n) = \sum\left[x(k) \times w(n-k)\right] \tag{3}$$

where k represents the index range of the convolution. Each sample in signal x(n) is weighted average with the corresponding part of the window function to achieve smoothing effect. We find that directly applying threshold comparison to detect peak points in envelope data is easily affected by false peaks caused by noise. To improve +P, a Hilbert filter is applied to convert peak points to zero crossing points, effectively reducing the sensitivity of noise peaks. In this paper, we change the standard filtering coefficients to reduce computational complexity, as follows:

$$[-1/25, \cdots, 0, -1/2, 0, -1, 0, 1, 0, 1/2, 0, \cdots, 1/25] \tag{4}$$

To achieve a Hilbert transform time length of 1 s, four zero values are inserted between every two non-zero coefficients, and zero-skipping method is applied to reduce computing power. The convolution of filtering is the same as (3). Then, zero crossing points are detected for the output results according to the following two rules: 1) Adaptive slope threshold comparison. Set the threshold Vth = Slope$_n$/8, where Slope$_n$ is the difference between the two values before and after n-th zero crossing. Every time a new zero crossing point is detected, if it is greater than Vth, the point is considered a candidate R-peak. Regardless of whether the current zero crossing meets the threshold, Vth will be updated for the next detection. 2) Prohibition period for detection. Zero crossing is no longer detected within 0.2 s after each detection, which can reduce the interference of pseudo zero crossings. Combining the Hilbert transform and the proposed zero crossing detection method, +P is improved.

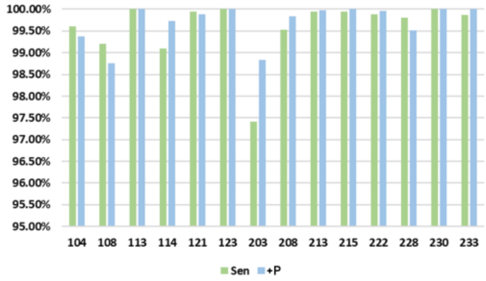


Fig. 4. The evaluation results of the proposed R-peak detection method.

Finally, R-peak verification is performed. At the initial stage, the standard deviation (SD) for the 8 detected RRI is calculated. After meeting the SD, the average value (AVG) will be calculated. If the change in the new RRI relative to AVG does not exceed 25%, the current detected R-peak is considered reliable. However, the large insertion noise between R-peaks can reduce RRI and lead to missed detections. To ensure high Sen, cumulative bias verification is adopted. Specifically, the small RRI will be temporarily saved and added to the next RRI for comparison with the AVG to further validate whether the current R peak is valid. The performance of the proposed detection method is evaluated using all 48 records in the MIT-BIH arrhythmia database [6], the true positives (TP), false positives (FP), and false negatives (FN) of each record are obtained, and the Sen and +P are calculated as follows:

$$Sen = TP/\left(TP + FN\right),\ +P = TP/\left(TP + FP\right) \tag{5}$$

Fig. 4 shows the evaluation results of part of ECG records. The total Sen and +P of all records are 99.78% and

99.74%, respectively. We compared the performance of the proposed method with other R-peak detection methods used for hardware, as shown in Table I. It can be observed that our proposed method realizes better Sen and +P.

TABLE I. PERFORMANCE COMPARISONS WITH OTHER METHODS.

Ref.	[7]	[8]	[9]	This Method
Method	deriv. & squaring	wavelet transform	slope comparison	Hilbert & zero crossing detection
Sen (%)	95.65	99.72	96.9	99.78
+P (%)	99.36	99.49	99.1	99.74

C. Data Transfer Module

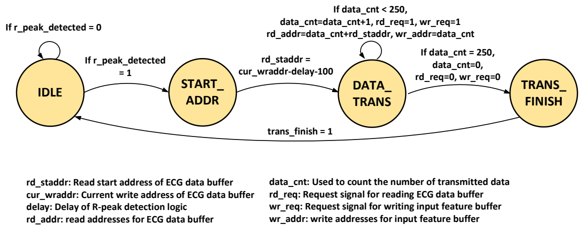

If r_peak_detected = 0
IDLE
If r_peak_detected = 1
START_ ADDR
rd_staddr = cur_wraddr-delay-100
If data_cnt < 250,
data_cnt=data_cnt+1, rd_req=1, wr_req=1
rd_addr=data_cnt+rd_staddr, wr_addr=data_cnt
DATA_ TRANS
If data_cnt = 250,
data_cnt=0,
rd_req=0, wr_req=0
TRANS_ FINISH
trans_finish = 1
rd_staddr: Read start address of ECG data buffer
cur_wraddr: Current write address of ECG data buffer
delay: Delay of R-peak detection logic
rd_addr: read addresses for ECG data buffer
data_cnt: Used to count the number of transmitted data
rd_req: Request signal for reading ECG data buffer
wr_req: Request signal for writing input feature buffer
wr_addr: write addresses for input feature buffer

Fig. 5. The workflow of the data transfer module.

The data transfer module is used to send the 100 data points before the R-peak and 150 data points after the R-peak from ECG data buffer to the input feature buffer for the classification. Fig. 5 illustrates the workflow of the data transfer module. The starting address (rd_staddr) for data transferring is located at 100 data points before the R-peak index. Once an R-peak is detected, rd_staddr is calculated according to the write address of the current input data (cur_wraddr). Note that the delay caused by FIR filtering in R-peak detection needs to be subtracted. After rd_staddr is determined, the data transferring begins. The data_cnt signal is used to count the number of transmitted data points, which is also used as an offset to add to rd_staddr for generating read addresses (rd_addr). The addresses are input to the ECG data buffer along with the read request signal (rd_req). In addition, the data_cnt signal can be regarded as the write addresses (wr_addr), which are fed to the Input feature buffer with the write request signal (wr_req). After the data_cnt reaches 250, the transfer finish signal is set to 1.

D. Classification Engine

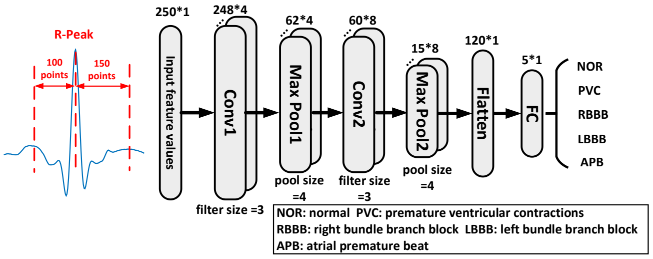


Fig. 6. The proposed 1-D CNN model for arrhythmia detection.

The proposed classification engine is based on 1-D CNN. In this paper, an algorithm model for detecting arrhythmia is constructed for dataflow analysis and functional verification of the classification engine, as shown in Fig. 6. It contains 2 convolutional layers, 2 pooling layers and 1 fully connected layer. The input feature size is 1×250. MIT-BIH arrhythmia database is used to train the model. Twenty thousand segmented heartbeats centered around the R-peak with a length of 250 are obtained from the database through a preprocessing algorithm. The training set and testing set each contains 10000 ECG records, and the proportion of each arrhythmia type is roughly the same.

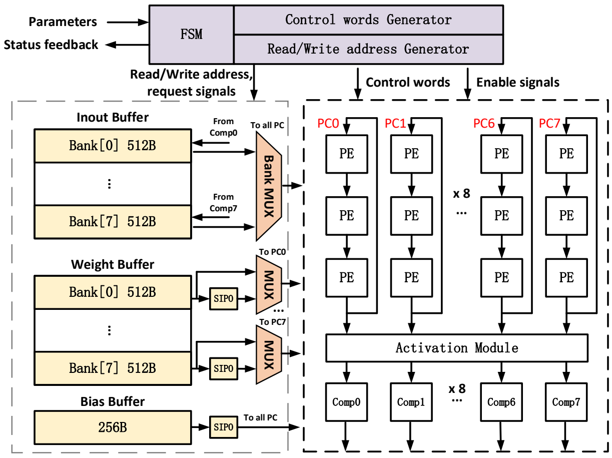


Fig. 7. Overall architecture of the classification engine.

TABLE II. PARAMETERS OF DATA BUFFERS.

Buffer	Bit Width (bits)	Storage Type	Capacity (Bytes)	Bank Number
Image Buffer	16	Dual Port SRAM	512	8
Weight Buffer	16	Single Port SRAM	512	8
Bias Buffer	16	Single Port SRAM	256	1

The architecture is shown in Fig. 7. The input feature values, weights and biases are stored in Inout Buffer, Weight Buffer and Bias Buffer, respectively. The parameters of all data buffers are shown in Table II. The Inout Buffer and Weight Buffer all contains 8 banks, each bank transmits data to the corresponding PE chain (PC).

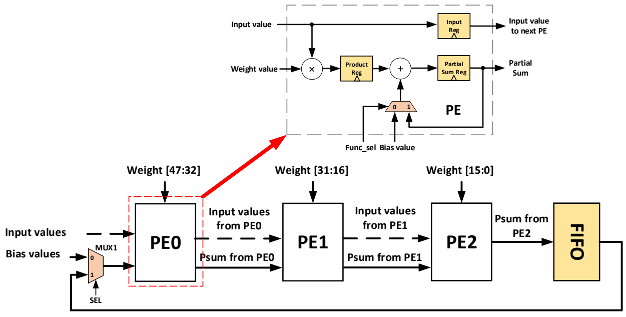
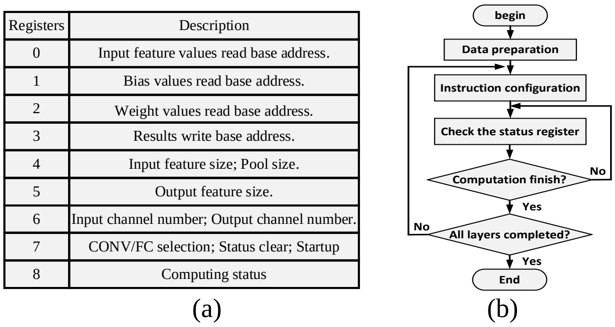


Fig. 8. A PE chain composed of 3 PEs.

To obtain high throughput, a PE array consisting of 8 PCs is applied, with each PC containing 3 PEs, as shown in Fig. 8. Each PE has multiplication and addition function, which completes the computation of one corresponding weight. The output partial sum (Psum) is directly passed to next PE as the addend without additional storage. Thus, parallelism within filters for convolution with a filter size = 3 can be performed through the pipeline in PC. Simultaneously, multiple output channels (or filters) are computed in parallel across the PCs. The output results are temporarily stored in a first-in-first-out (FIFO). The time-multiplexing method is used in convolution and fully connected computing. For convolution with multiple input channels, the results of the previous input channel in the FIFO will be read as the bias values for the next input channel, so all input channels can be completed in one PC. For fully connected computing, all multiply-accumulate operation can be performed by reusing the first PE in the PC. This method effectively reduces the cost of chip area and overall power. After the computing on PCs is completed, the results are read out from the FIFO and fed into the activation module, which performs the ReLU function by comparing the sign bits of input data. Finally, the maximum pooling is completed using comparators. The pooling results are stored in the corresponding bank in the Inout Buffer as input features for the next convolutional layer.

Date reuse strategy is applied to reduce the memory access power. The input feature values are broadcasted to all PCs and cached in registers in PE, and will be passed directly to the next PE. The serial-in-parallel-out (SIPO) circuits are used to update the weight values for one input channel to a 48-bit register, which does not to be changed until current channel is calculated, effectively reducing the frequency of accessing SRAM and improving energy efficiency.

E. Central Controller



Registers	Description
0	Input feature values read base address.
1	Bias values read base address.
2	Weight values read base address.
3	Results write base address.
4	Input feature size; Pool size.
5	Output feature size.
6	Input channel number; Output channel number.
7	CONV/FC selection; Status clear; Startup
8	Computing status

Fig. 9. (a) The description of parameter registers. (b) The flow of processing 1-D CNN.

The central controller consists of instruction and status registers, decoder/encoder, and finite state machine (FSM), responsible for generating control words for the PCs and controlling access to data buffers. The detailed description of registers is shown in Fig. 9(a). 1-D CNN is processed layer by layer. After one convolutional layer and pooling layer are computed, the computation finish flag is generated, then the main processor detects this flag and configures the instructions for the next computation, as shown in Fig. 9(b).

III. Implementation Results

The proposed ECG coprocessor is implemented using UMC 55 nm Low-K CMOS technology. The post placement and routing are completed using the Innovus Implementation System. Fig. 10(a) shows the chip layout. The core size is 0.94 mm × 0.94 mm, and the entire chip size is 1.18 mm × 1.18 mm. The coprocessor is post simulated, the average power consumption is 17.6 mW at 100 MHz clock frequency with a power supply of 1.08 V, the details of the power analysis are shown in Fig. 10(b). Each classification latency is 12.25 μs, and a throughput of 4.80 GMAC/s is achieved. We compared the performance with other ECG processors, as shown in Table III. It can be seen that our design has higher throughput and better energy efficiency, while supporting flexible instruction configurations.

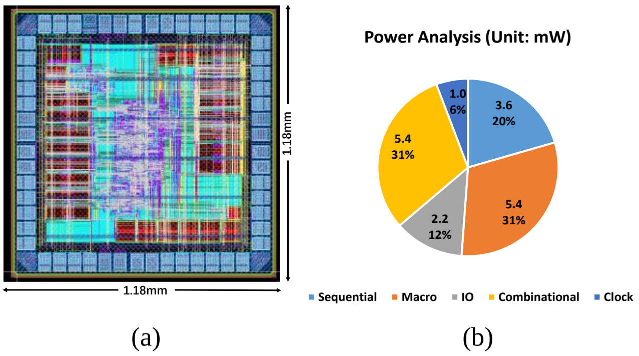


Fig. 10. (a) The layout photograph. (b) The details of power analysis.

TABLE III. Performance Summary and Comparison.

	[3]	[4]	[5]	[10]	[11]	This Work
Process (nm)	65	40	40	65	40	55
Method	ANN	1-D CNN	ELM	RC-NN	LC-SNN	1-D CNN
Supply Voltage (V)	0.75	1.1	0.9	1.2	1.1	1.08
Frequency (Hz)	2.5M	2M	70M	40K	100M	100M
Area (mm^2)	1.74	2.044	0.212	0.24	0.65	1.39
Power (mW)	0.087	-	19.2	-	-	17.6
Throughput (GMAC/s)	-	0.096	3.45	0.007	-	4.80
Power EFF (GMAC/s/W)	-	271	180.1	80	-	272.7
Energy EFF (nJ/Classifi.)	2250	3930	477.3	-	750	215.6
Classification Types	2	5	2	2	5	5[a]
Accuracy (%)	99.3	98.9	92	86.8	98.2	97.7
Instruction Configuration	No	Yes	No	No	No	Yes

[a] To meet the AAMI standard [12], a new 1-D CNN can be configured on this coprocessor.

In [3], [5], and [10], ECG can only be classified into 2-class. The complex FIR engine and NN engine in [3] result in high energy consumption. The PCA transformation and ELM in [5] require 52KB of memory, the intensive computations lead to high power. In [10], the NN based on analog circuits makes the accuracy sensitive to circuit noise and temperature variations. In [4], the working modes of the convolutional and fully connected layers have idle PEs, leading to extra chip area cost and relatively low energy efficiency. In [11], the 1,024 neurons-SNN has a storage overhead of up to 69KB, and the long classification latency consumes more energy. Except for [4], other works do not support instruction configuration, which makes them unable to flexibly adapt to different classification scenarios.

IV. Conclusion

In this paper, an energy-efficient ECG coprocessor is proposed. The proposed R-peak detection engine has high Sen and +P but low hardware cost, and the configurable classification engine can effectively perform computations of different 1-D CNNs for ECG classification. The hardware implementation results show that the proposed coprocessor achieves flexible and energy efficient on-chip local ECG anomaly analysis. In the future, we will optimize power by using low bit-width weights and integrate the proposed coprocessor with the ECG acquisition front-end.

REFERENCES

[1] Y. Zhao, Z. Shang, and Y. Lian, “A 13.34 μW Event-Driven Patient-Specific ANN Cardiac Arrhythmia Classifier for Wearable ECG Sensors,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 2, pp. 186–197, Apr 2020.

[2] M. Janveja, R. Parmar, M. Tantuway, and G. Trivedi, “A DNN-Based Low Power ECG Co-Processor Architecture to Classify Cardiac Arrhythmia for Wearable Devices,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 4, pp. 2281–2285, Apr 2022.

[3] J. Liu, Z. Zhu, Y. Zhou, N. Wang and G. Dai, “4.5 BioAIP: A Reconfigurable Biomedical AI Processor with Adaptive Learning for Versatile Intelligent Health Monitoring,” in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, pp. 62–64, Feb 2021.

[4] J. Lu, D. Liu, X. Cheng, L. Wei, A. Hu, and X. Zou, “An Efficient Unstructured Sparse Convolutional Neural Network Accelerator for Wearable ECG Classification Device,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 11, Art. no. 11, Nov 2022.

[5] Y.C. Chuang, Y.T. Chen, H.T. Li, and A.Y.A. Wu, “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.

[6] P. PhysioToolkit, “Physionet: Components of a new research resource for complex physiologic signals”, *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000.

[7] H.M. Wang, Y.L. Lai, M.C. Hou, S.H. Lin, and B.S. Yen, “A ±6ms-accuracy, 0.68mm2 and 2.21μW QRS detection ASIC,” in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 1372–1375, May 2010.

[8] Y. Zou, J. Han, S. Xuan, S. Huang, and X. Weng, “An Energy-Efficient Design for ECG Recording and R-Peak Detection Based on Wavelet Transform,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 2, Feb 2015.

[9] Y. Yin, S. M. Abubakar, S. Tan, and H. Jiang, “A 17.7-pJ/Cycle ECG Processor for Arrhythmia Detection with High Immunity to Power Line Interference and Baseline Drift,” in *2020 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 1–4, Nov 2020.

[10] S. Tannirkulam Chandrasekaran and S. Prashant Bhanushali, “Toward Real-Time, At-Home Patient Health Monitoring Using Reservoir Computing CMOS IC,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 4, Dec 2021.

[11] H. Chu, Y. Yan, L. Gan, H. Jia and L. Qian, “A Neuromorphic Processing System With Spike-Driven SNN Processor for Wearable ECG Classification,” IEEE Transactions on Biomedical Circuits and Systems, vol. 16, no. 4, pp. 511–523, Aug. 2022.

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