ECG Classification using Binary CNN on RRAM Crossbar with Nonidealities-Aware Training, Readout Compensation and CWT Preprocessing

Tiancheng Cao[1,2], Zhongyi Zhang[1,2], Wang Ling Goh[1], Chen Liu[2], Yao Zhu[2], Yuan Gao[2]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Republic of Singapore
[2]Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Republic of Singapore
Email: CAOT0006@e.ntu.edu.sg, gaoy@ime.a-star.edu.sg

***Abstract*—This paper presents an electrocardiogram (ECG) signal classification method using binary CNN implemented on RRAM crossbar arrays. A new RRAM crossbar structure is proposed to provide input-dependent references for adaptive readout quantization. Hence, binary weights can be represented with just one RRAM cell, instead of the conventional differential cell, leading to the reduction of crossbar size by half. Furthermore, the impacts of crossbar nonidealities is mitigated with nonidealities-aware training and in-situ readout compensation. On the other hand, bandpass filter (BPF) based continuous wavelet transform (CWT) approximation is applied for ECG signal preprocessing to enhance the feature extraction. Implemented on 64×64 binary RRAM crossbar arrays, the proposed binary CNN achieved 98.9% classification sensitivity on 10-class MIT-BIH ECG datasets, with only 0.7% sensitivity drop compared to the software version with FP32 weights.**

Keywords—binary neural network, processing-in-memory, ECG classification, neuromorphic system

I. Introduction

In recent years, convolutional neural network (CNN) has achieved impressive results in various classification tasks such as image recognition [1-4] and language processing [5]. CNN has also shown great potential in biomedical signal processing, particularly in the analysis of electrocardiogram (ECG) signal for cardiovascular diseases (CVD) [6]. Wearable ECG monitoring devices offer critical information for the early detection of heart arrhythmias and heart failure [7]. So that early clinical intervention can be triggered. There has been an increasing interest in integrating artificial intelligence function into ECG devices to achieve intelligence at the edge. However, implementing CNN on resource constraint wearable devices poses great design challenges, due to the limited memory size and computation capacity in the wearable device [8]. Binary neural network (BNN), in which weights are quantized to only binary values [9] is considered as a promising alternative to CNN. BNN offers significant reduction in memory size while maintaining comparable classification accuracy [10]. The objective is to strike a balance between model complexity, inference accuracy, computing resources, and power consumption [11-12].

RRAM crossbar-based neural network presents an appealing solution for high energy efficiency [13]. A few RRAM crossbar-based neural networks have been reported for ECG signal classification [14-16]. However, most prior designs still rely on a differential pair cell structure to represent one bit due to the challenge to generate negative conductance. Recent study introduced a one device one cell scheme, which reduces the required crossbar hardware by half [17]. However, this structure has potential readout error risk when the output current approaches zero. Additionally, fabricating RRAM array with high uniformity still faces great challenges [18-21]. Non-ideal factors such as resistance variation and parasitic line resistance significantly impact the accuracy of crossbar arrays [22-23]. In summary, implementing a robust BNN on RRAM crossbar array remains a major challenge.

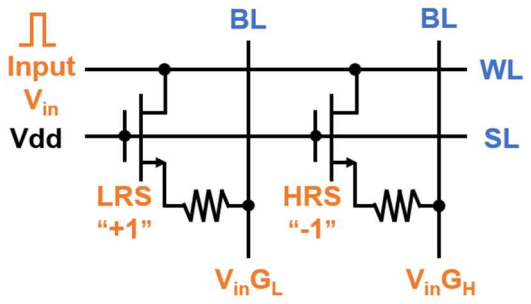
This paper introduces a an electrocardiogram (ECG) signal classification method using binary CNN implemented on RRAM crossbar arrays. Various techniques are proposed to improve the BNN performance by mitigating RRAM crossbar nonidealities and enhancing the ECG feature extraction through the preprocessing of CNN inputs. Two auxiliary columns of RRAM in low resistance level (LRS) and high resistance level (HRS) are added into crossbar structure to generate input-dependent references for the readout circuit. These references help to compensate the offsets that may introduce errors during readout. To mitigate the impact of crossbar nonidealities, they are incorporated into the BNN training procedure. Furthermore, a bandpass filter (BPF) based continuous wavelet transform (CWT) approximation is proposed to preprocess the ECG data to enhance the feature extraction. Simulation results showed that the proposed binary CNN achieved classification sensitivity of 98.89% for the MIT-BIH ECG dataset [16] specified in Table.I on 64×64 binary RRAM crossbar arrays. This paper is organized as follows: Section II introduces the proposed in-situ readout compensation method. Section III elaborates on the implementation of the binary convolutional neural network and the nonidealities-aware training method. Section IV presents the simulation results using the ECG dataset. Finally, Section V concludes the paper.

II. In-situ BNN Readout Compensation

A. Binary Cell Representation

To represent negative values in the weight matrix, the conventional approach is to utilize a differential pair of RRAM crossbars, with one for positive values and another for negative values [24-25]. However, in the context of Binary Neural Networks (BNNs), where each cell requires only 1 bit, an alternative solution is feasible. This solution involves using a single device to represent one cell,

This work was supported by the Agency for Science, Technology and research (A*STAR), Singapore under the Nanosystems at the Edge programme, grant no. A18A1b0055 and the Cyber-Physiochemical Interface programme, grant No. A18A1b0045.



Input	RRAM	Weight	BL Current	I_{ref}	Cell Output
V_{in}	LRS	+1	$V_{in}G_L$	$0.5V_{in}(G_L+G_H)$	$0.5V_{in}(G_L-G_H)$
V_{in}	HRS	-1	$V_{in}G_H$	$0.5V_{in}(G_L+G_H)$	$-0.5V_{in}(G_L-G_H)$

Fig. 1. Methodology of the proposed binary RRAM cell. I_{ref} comes from the extra auxiliary columns which is not shown. Two cells represent different situation for different RRAM states and different input signals.

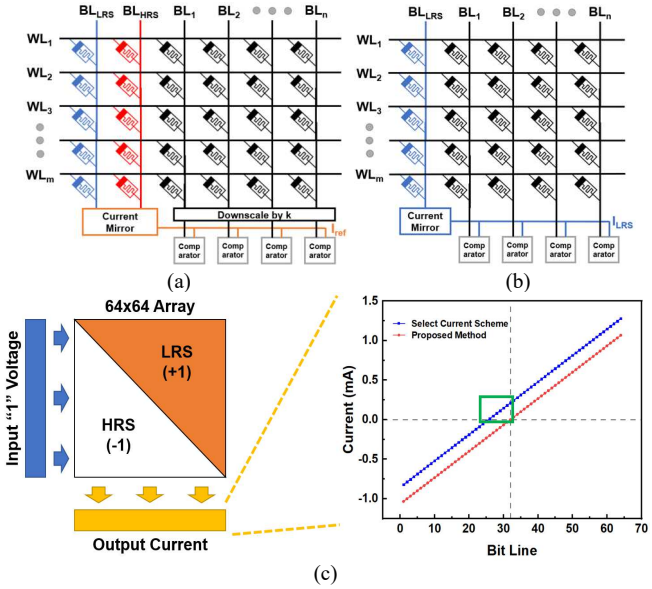


Fig. 2. (a) Crossbar structure of the proposed one-device-one-cell dynamic weight mapping method with two extra columns with fixed LRS/HRS states. (b) The one-device-one-cell Select Column Scheme presented in previous work [14]. (c) Comparison of bit line output current using Select Current Scheme and the proposed method.

commonly referred to as a 1T1R cell. Implementing 1T1R cells reduces both the area and energy consumption by half [17].

Two binary RRAM cell is shown in Fig. 1. The input is represented by voltage signals on the word line. Two levels of the RRAM, low resistance level (LRS) and high resistance level (HRS), indicate binary weight value "+1" and "-1" respectively. When the word line receives a voltage pulse, V_{in}, the current flows through the RRAM. If RRAM is at LRS state, current, $V_{in}G_L$, will flow through the bit line, where G_L is the conductance of the LRS RRAM. Similarly, if RRAM is at HRS state, current, $V_{in}G_H$, will flow through the bit line, where G_H is the conductance of the HRS RRAM.

B. In-situ Readout Compensation

The schematic of the proposed in-situ readout compensation is illustrated in Fig. 2(a) using a 64×66 crossbar array. In this configuration, the leftmost two columns of the

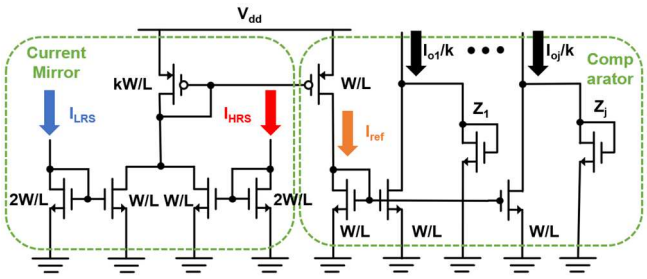


Fig. 3. The diagram of the current mirror block and the comparator block.

crossbar serve as additional auxiliary reference current sources. One column consists of all cells set to low resistance level (LRS), while the other column is set to high resistance level (HRS). Since the input voltage vector is also applied to these two columns, the resulting output reference current becomes input-dependent. This represents a significant distinction compared to using fixed external references.

When the input voltage is pumped into the word lines, the accumulation cell output current of the LRS and HRS columns, I_{LRS} and I_{HRS}, which is dependent on the input, will flow into the current mirror block. The diagram of the current mirror block is given in Fig.3. The output of the current mirror block serves as the reference current

$$I_{ref} = \frac{1}{2k}(I_{LRS} + I_{HRS}) \quad (1)$$

where k is the downscale coefficient to reduce the power consumption of the module.

The reference current is then flowed into the comparator with the downscaled output current from j[th] bit line, I_{Oj}/k. The comparator circuit is shown in Fig.3. The output can be summarized as

$$Z_j = \frac{I_{Oj}}{k} - I_{ref} \quad (2)$$

where Z_j is the j^{th} column output neuron of the current neural network layer.

The block diagram of the previously reported 1T1R readout scheme [17] is shown in Fig. 2 (b). This scheme uses one auxiliary column as reference and set all cells to LRS. The current generated from this selected column, I_{sel}, will be divided by two then compared to the output current from the other bit lines. However, this selected reference current neglect the currents generated from HRS cell, which will accumulate and cause the error. To compare the performance of the proposed method with Select Column Scheme, an input vector and array weights test pattern is constructed as shown in Fig. 2(c). All input voltage is set as "1", and the cell state is set either to LRS or HRS diagonally as shown in Fig. 2(c). The simulated output current of each column is shown. The result obtained with the proposed method is represented by the red line which fit the BNN calculation well. When there are more HRS cells, the output current is negative. When there are more LRS cell, the output current is positive. In contrast, the result of select column scheme which is shown in blue color generates errors in this test pattern. The green box region, where the number of LRS cell is slightly less than the number of HRS cell is supposed to generate negative

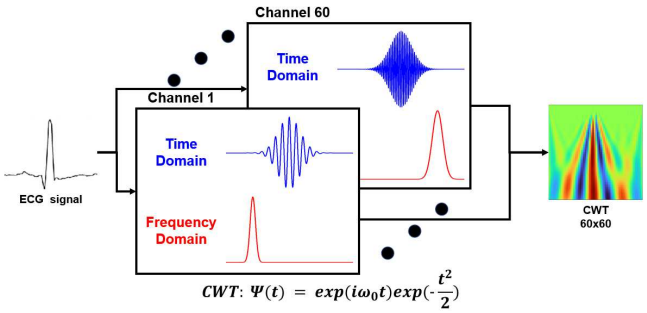


Fig. 4. BPF-based 60-channel CWT-approximation signal processing of ECG in this work.

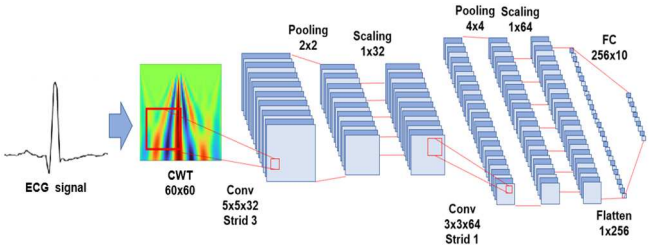


Fig. 5. The structure of the binary convolutional neural network used in this work for training and test.

Table I. 10-class MIT-HIB ECG dataset

AAMI Class	MIT-BIH Heartbeat Types
Normal beats	Normal beats (N), Left bundle branch block (L), Right bundle branch block (R), Atrial escape beat (a)
Supraventricular ectopic Beats	Atrial premature contraction (A), Junctional premature beat (J), Non-conducted P-wave (x)
Ventricular ectopic beats	Ventricular escape beat (E)
Fusion beats	Fusion of ventricular and normal beat (F)
Unclassifiable beats	Unclassified beat (Q)

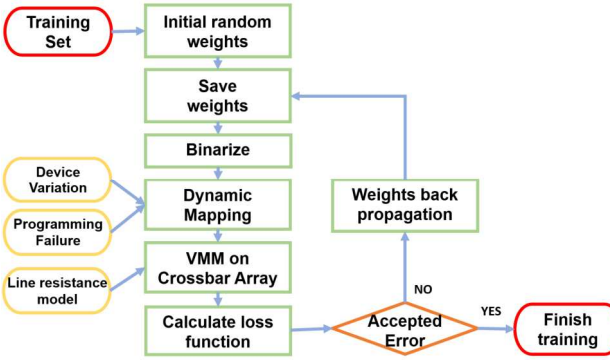


Fig. 6. Flow chart of binary nonidealities aware training.

output but show wrong positive value due to neglection of accumulated current of HRS RRAM.

III. Binary CNN Network with CWT Processing

A. Data Preprocessing

The original one-dimensional (1D) ECG signal is firstly converted to 2D time-frequency feature by applying continuous wavelet transform (CWT). Compared with short-time Fourier transform (STFT), CWT provides better time-frequency resolution, and it is able to capture non-stationary characteristics of ECG signals. The employed CWT has scales from 1 to 60 and Morlet wavelet is used as mother wavelet with sampling frequency of 360Hz. The Transform is implemented with band pass filter (BPF) as shown in Fig.4. BPF with a high Q-factor and high order can approximate the high frequency resolution of CWT. Additionally, the time-domain resolution of CWT can be accomplished via additional time window steps. This BPF-based approximation technique offers an advantage over traditional CWT signal processing [26] by facilitating more convenient adjustment of time-frequency characteristics.

B. Neural Network Structure

Fig. 5 illustrates the structure of the binary CNN network. The network comprises 2 convolutional layers, 2 pooling layers, 2 channel scaling layers, and 1 fully-connected (FC) layer, totaling 21.5k parameters. In contrast to conventional CNN that employ batch normalization, the output of batch normalization in this network includes a scale and shift operation for each channel. Since the bias of the convolution layer in each block already serves as the shift function, an additional simple scaling operation is introduced to each channel. Through software training with floating point weights, a test sensitivity of 99.6% is achieved.

C. Binary Convolutional Layer

The convolutional layer is responsible for feature extraction, which is a critical step for binary CNN. Referring to the previous work that applied the convolution operation on multi-level RRAM-based crossbar [27], the binary convolution operation is designed based on the process reported in [28]. Both the binary kernels and the feature maps from the input can be unfolded. The kernel is reshaped into a column on the crossbar array. The reshaped kernels are appended column by column to form a multiple-kernel matrix. Similarly, the feature maps of next layer are created after the folding.

D. Nonidealities Modeling and Aware Training

The conductance of RRAM follows the normal distribution because memristors typically have a conductance variation [29]. Thus, the random normal distribution mask is applied to HRS and LRS level in the model as the variation. The impact of line resistance to crossbar-based system been explored in previous work [22]. It is negligible in relatively small size crossbars. However, as the array size grows, the line resistance effects become more severe, resulting in significant degradation. In this work, a highly efficient line resistance impact estimation model [30] is applied in the training and inference process, which compensates the error accumulation and voltage headroom when turning on the array cells simultaneously. With the line resistance, the voltage across the cell is reduced, the effective conductance also decrease. As a result, the currents collected from the bit lines are smaller than expected and can further reduce NN inference accuracy. The programming failure is simulated by randomly turning off the RRAM cells in crossbar simulation. Considering the above nonidealities, the nonidealities aware training shown in Fig. 6 is conducted during the system training.

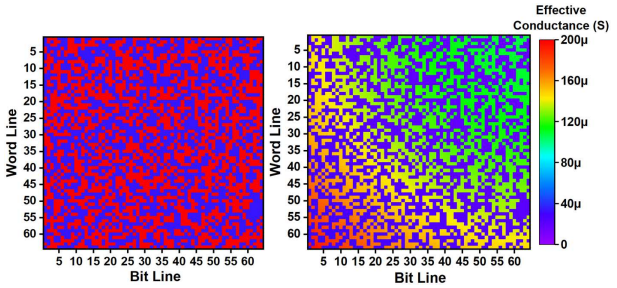


Fig. 7. The effective conductance of the 64×64 random weights crossbar array (a) without line resistance and (b) with 1.5Ω line resistance.

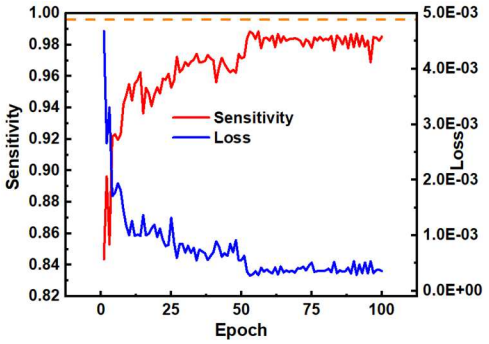


Fig. 8. Test inference sensitivity with nonidealities incorporated training under crossbar array nonidealities.

IV. System Evaluation

A. Nonidealities Evaluation

An efficient line resistance modeling for RRAM crossbar array considering the device variation is employed [30]. The impact of 1.5Ω line resistance on a random programed 64×64 crossbar array with 1-bit RRAM devices is shown in Fig. 7 by the effective cell conductance which is defined as the current collected from the j^{th} bit line from the (i, j) cell divided by the i^{th} row input voltage through the (i, j) cell as below.

$$effective\ G_{i,j} = \frac{I_{Oij}}{V_{ini}} \tag{3}$$

B. Binary CNN Evaluation

The binary CNN model shown in Fig. 5 is implemented in Python. The nonidealities parameters are based on the measured TaOx RRAM 64×64 crossbar array model [31], they are listed in Table. II. Additionally, due to the extreme imbalance in the distribution of the dataset samples, a random sampling approach is employed during training for the more frequent classes, while the less frequent classes are augmented to increase their sample size, thus ensuring equal representation of each class per epoch. For the same reason, sensitivity is used to evaluate the system instead of accuracy as the focus should be on identifying the minority of abnormal ECG signals rather than the majority of normal signals. Sensitivity is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN) as below:

$$sensitivity = \frac{TP}{TP+FN} \tag{4}$$

By taking the proposed readout compensation and nonidealities into the aware training with Pytorch, the inference sensitivity of the model can be trained to 98.89% as shown in Fig. 8, where the sensitivity of the software model is given as the reference line. The result has only 0.7% degradation as compared to the software model. Table. III summarize the performance of the proposed method and compared on MIT-HIB dataset with other recently published BNN implemented by memristor based crossbar. It can be observed that the proposed mapping and training offer the most comprehensive analysis. With only 1T1R cell and 21.5k parameters, a higher inference sensitivity is achieved.

Table II. Summary of nonidealities

Nonideality Factor	Value
Average R (kΩ)	12.9/33.8
Device Variation	25%
Array Size	64×64
Line Resistance	0.72 Ω
Failure	2%

Table III. Performance Comparison Table

	[14]	[16]	[32]	[33]	This Work
ECG class	2	15	8	5	10
Preprocessing	binary image	segment	STFT	segment	CWT
Input	16x20	N.R.	256x256	130	60x60
Architecture	2D CNN	FC+TCN +LSTM	2D CNN	1D CNN	2D CNN
Database balance	No	No	No	Yes	Yes
RRAM CIM	No	Yes	No	No	Yes
Neural Network	1Conv +2FC	Hierarchical	3FC	2Conv +1FC	2Conv +1FC
# of parameters	64k	N.R.	3.23M	N.R.	21.5k
Binary	0, 1	No	+1, -1	No	+1, -1
Haedware implementation	Yes	Yes	No	No	Yes
Nonideality aware	No	No	No	No	Yes
Accuracy/ Sensitivity	85.70%	98.29%	99.11%	97.50%	98.89%

N.R.: Not reported

V. Conclusion

In this paper, a nonidealities-aware in-situ readout compensation and network training scheme for binary convolutional neural network implemented on memristor crossbar arrays is presented. Auxiliary columns of RRAM with fixed LRS and HRS states are used to provide input-dependent references for adaptive readout binary quantization. The influence of crossbar array nonidealities on binary computation is also analyzed. The proposed readout compensation and training methods are demonstrated using a CNN implemented in PyTorch on a 64×64 binary RRAM crossbar array model for 10-class MIT-HIB ECG recognition, employing BPF-based 60-channel CWT approximation signal processing. The classification sensitivity achieved is 98.89%, which is only 0.7% lower than the result obtained from an ideal software model with FP32 weights.

References

[1] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[3] Y. Lu, V. L. Le and T. T. -H. Kim, "9.7 A 184 μ W real-time hand-gesture recognition system with hybrid tiny classifiers for smart wearable devices," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, pp. 156-158, 2021.

[4] Y. Lu, V. L. Le and T. T. -H. Kim, "A 184-μW error-tolerant real-time hand gesture recognition system with hybrid tiny classifiers utilizing edge CNN," *IEEE J. Solid-State Circuits*, vol. 58, no. 2, pp. 530-542, Feb. 2023.

[5] X. He, K. Wang, H. Huang, T. Miyazaki, Y. Wang, and S. Guo, "Green resource allocation based on deep reinforcement learning in content-centric IoT," *IEEE Trans. Emerg. Topics Comput.*, Feb. 13, 2018.

[6] S. S. Virani et al., "Heart disease and stroke statistics—2021 update," *Circulation*, vol. 143, no. 8, pp. 1–6, Feb. 2021.

[7] T. Yang, L. Yu, Q. Jin, L. Wu, and B. He, "Localization of origins of premature ventricular contraction by means of convolutional neural network from 12-Lead ECG," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1662–1671, Jul. 2018.

[8] X. Sun, R. Liu, X. Peng and S. Yu, "Computing-in-Memory with SRAM and RRAM for Binary Neural Networks," *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2018.

[9] M. Courbariaux, et al., "Binarized neural network: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv: 1602.02830, 2016.

[10] M. Rastegari, et al., "XNOR-net: ImageNet classification using binary convolutional neural networks," arXiv: 1603.05279, 2016.

[11] Y. Zhao, Z. Shang, and Y. Lian, "A 13.34 μw event-driven patient-specific ANN cardiac arrhythmia classifier for wearable ECG sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 186–197, Apr. 2020.

[12] M. Saeed et al., "Evaluation of level-crossing ADCs for event-driven ECG classification," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 6, pp. 1129–1139, Dec. 2021.

[13] S. Agarwal, O. D. Parekh, T. Quach, C. D. James, J. B. Aimone and M. Marinella, "The energy scaling advantages of RRAM crossbars," *2015 Fourth Berkeley Symposium on Energy Efficient Electronic Systems (E3S)*, Berkeley, CA, pp. 1-3, 2015.

[14] D. L. T. Wong, Y. Li, D. John, W. K. Ho and C. H. Heng, "Resource and energy efficient implementation of ECG classifier using binarized CNN for edge AI devices," *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, Daegu, Korea, pp. 1-5, 2021.

[15] E. Esmanhotto et al., "High-density 3D monolithically integrated multiple 1T1R multi-level-cell for neural networks," *2020 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2020.

[16] S. Diware et al., "Severity-based hierarchical ECG classification using neural networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 1, pp. 77-91, Feb. 2023.

[17] Y. -F. Qin, R. Kuang, X. -D. Huang, Y. Li, J. Chen and X. -S. Miao, "Design of high robustness BNN inference accelerator based on binary memristors," in *IEEE Trans. Electron Devices*, vol. 67, no. 8, pp. 3435-3441, Aug. 2020.

[18] Y. Jeong and W. Lu, "Neuromorphic computing using memristor crossbar networks: a focus on bio-inspired approaches," in *IEEE Nanatechnol. Mag.*, vol. 12, no. 3, pp. 6-18, Sept. 2018.

[19] Z. Dong *et al*., "Convolutional neural networks based on RRAM devices for image recognition and online learning tasks," in *IEEE Trans. Electron Devices*, vol. 66, no. 1, pp. 793-801, Jan. 2019.

[20] M. Prezioso, F. Merrikh-Bayat, B. Hoskins *et al*, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61–64, 2015.

[21] L. Xu, C. Li and L. Chen, "Analog memristor based neuromorphic crossbar circuit for image recognition," *2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*, Wuhan, pp. 155-160, 2015.

[22] M. E. Fouda, J. Lee, A. M. Eltawil and F. Kurdahi, "Overcoming crossbar nonidealities in binary neural networks through learning," *2018 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2018

[23] T. Hirtzlin et al., "Outstanding bit error tolerance of resistive RAM-based binarized neural networks," *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2019.

[24] X. Sun, X. Peng, P. Chen, R. Liu, J. Seo and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+1, −1) weights and (+1, 0) neurons," *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018.

[25] T. Cao, W. L. Goh and Y. Gao, "Performance analysis of convolutional neural network using multi-level memristor crossbar for edge computing," *2020 3rd International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2020.

[26] A. J. Casson and E. Rodriguez-Villegas, "A 60 pW gmC continuous wavelet transform circuit for portable EEG systems," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1406-1415, June 2011.

[27] X. Sun, X. Peng, P. Chen, R. Liu, J. Seo and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+1, −1) weights and (+1, 0) neurons," *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018.

[28] T. Cao et al., "RRAM-PoolFormer: a resistive memristor-based PoolFormer modeling and training framework for edge-AI applications," *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023.

[29] C. Li, M. Hu, Y. Li *et al*, "Analogue signal and image processing with large memristor crossbars," *Nat Electron,* vol 1, pp. 52–59, 2018.

[30] T. Cao, C. Liu, Y. Gao and W. L. Goh, "Parasitic-aware modeling and neural network training scheme for energy-efficient processing-in-memory with resistive crossbar array," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 12, no. 2, pp. 436-444, June 2022.

[31] W. Wang, et al, "Enabling neuromorphic computing: BEOL integration of CMOS RRAM chip and programmable performance," *2019 IEEE International System-on-Chip Conference (SOCC)*, Singapore, pp. 354-358, 2019.

[32] A. Ullah, S. M. Anwar, M. Bilal, and R. M. Mehmood, "Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation," *Remote Sens.*, vol. 12, no. 10, p. 1685, May 2020.

[33] D. Li, J. Zhang, Q. Zhang and X. Wei, "Classification of ECG signals based on 1D convolution neural network," *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Dalian, China, 2017.