Benchmarking RRAM Crossbar Arrays for Epileptic Seizure Prediction

Ahmedul Khan, Shiva Maleki Varnosfaderani, Mohammad Alhawari, Gozde Tutuncuoglu

Department of Electrical and Computer Engineering, Wayne State University, Detroit, USA
E-mail: {hn6828, shiva.maleki, alhawari, gozde}@wayne.edu

2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS) | 979-8-3503-8717-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/MWSCAS60917.2024.10658668

***Abstract*—This paper explores an energy-efficient resistive random access memory (RRAM) crossbar array framework for predicting epileptic seizures using the CHB-MIT electroencephalogram (EEG) dataset. RRAMs have significant potential for in-memory computing, offering a promising solution to overcome the limitations of the traditional Von Neumann architecture. By integrating a domain-specific feature extraction approach and evaluating the optimal RRAM hardware parameters using the NeuroSim+ benchmarking platform, we assess the performance of RRAM crossbars for predicting epileptic seizures. Our proposed workflow achieves accuracy levels above 80% despite the EEG data being quantized to 1-bit, highlighting the robustness and efficiency of our approach for epileptic seizure prediction.**

***Index Terms*—RRAM, epilepsy, EEG, benchmarking, in-memory computing**

I. Introduction

The growing power consumption of the state-of-the-art CMOS systems, coupled with continuously increasing computational load patterns, pose significant challenges to address complex machine learning problems without compromising the energy-efficiency and overall system performance. Such limitations cause bottlenecks in data processing and communication capabilities of edge computing infrastructure, which is critical for machine learning in healthcare applications, e.g. health monitoring systems via physiological signal processing. Enhancing the speed and energy efficiency of real-time physiological data analysis holds immense potential for advancing the integration of artificial intelligence (AI) technologies within medical Internet-of-Things (IoT) devices. Resistive Random Access Memory (RRAM) devices stand as a highly promising technology in this regard to enable energy-efficient in-memory computing for healthcare applications [1].

Advancements in computational methods and technology are particularly critical in the prediction and management of neurological disorders such as epilepsy. Epilepsy, characterized by its unpredictable seizure episodes, poses fundamental diagnostic and treatment challenges [2]. However, the integration of electroencephalogram (EEG) signal analysis into predictive models has demonstrated a strong potential to revolutionize epilepsy care. Machine learning algorithms employed on pre-processed EEG datasets can uncover subtle patterns and biomarkers that correlate with increased seizure risk [3], [4]. This methodology not only enhances the accuracy of seizure prediction but also facilitates the development of personalized medicine strategies, optimizing treatment regimens based on individual patient profiles.

In this paper, we explore an energy-efficient in-memory computing framework for RRAM crossbar arrays to predict epileptic seizures from the EEG data. In this study, we use the CHB-MIT dataset, which includes EEG data from 23 epileptic patients from the Children's Hospital of Boston that were captured at a sampling rate of 256 Hz [5], [6]. Following the EEG signal processing step, we use the NeuroSim+ benchmarking platform [7] to assess the performance of RRAM crossbars for predicting epileptic seizures.

This paper is organized as follows: Section II describes the EEG signal pre-processing and data processing steps. Section III details the NeuroSim+ benchmarking process as a function of device metrics, and Section IV concludes the findings.

II. EEG Signals and Preprocessing

The architecture of our suggested model is depicted in Figure 1. Our proposed workflow consists of the following: EEG data pre-processing, feature extraction, normalization and quantization of data processing steps, and two-layer fully connected feed-forward neural network (FCNN) classifier. Additional details about each step are provided in the following:

A. EEG Pre-processing

In CHB-MIT dataset, the number of channels varies between 23 and 26; however, the following 18 channels are consistently shared across all patients: FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FZ-CZ, CZ-PZ, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, and P8-O2. For our analysis, we only used four channels, FP1-F7, FP2-F8, P8-O2, and T7-P7, which covered a wider region, allowing for better spatial coverage of the brain. As a result, abnormalities or epileptic activity originating from different regions of the brain can be detected more accurately. After identifying the selected channels, each 60-minute raw scalp electroencephalography clip is split into 5-second non-overlapping segments.

Next, we extracted one-hour-long *preictal* and *interictal* samples. To extract preictal samples, we consider a five-minute gap from seizure onset. Two-hours gap before and after seizure onset is also considered to extract interictal samples. We label each preictal sample as 1 and each interictal sample as 0.

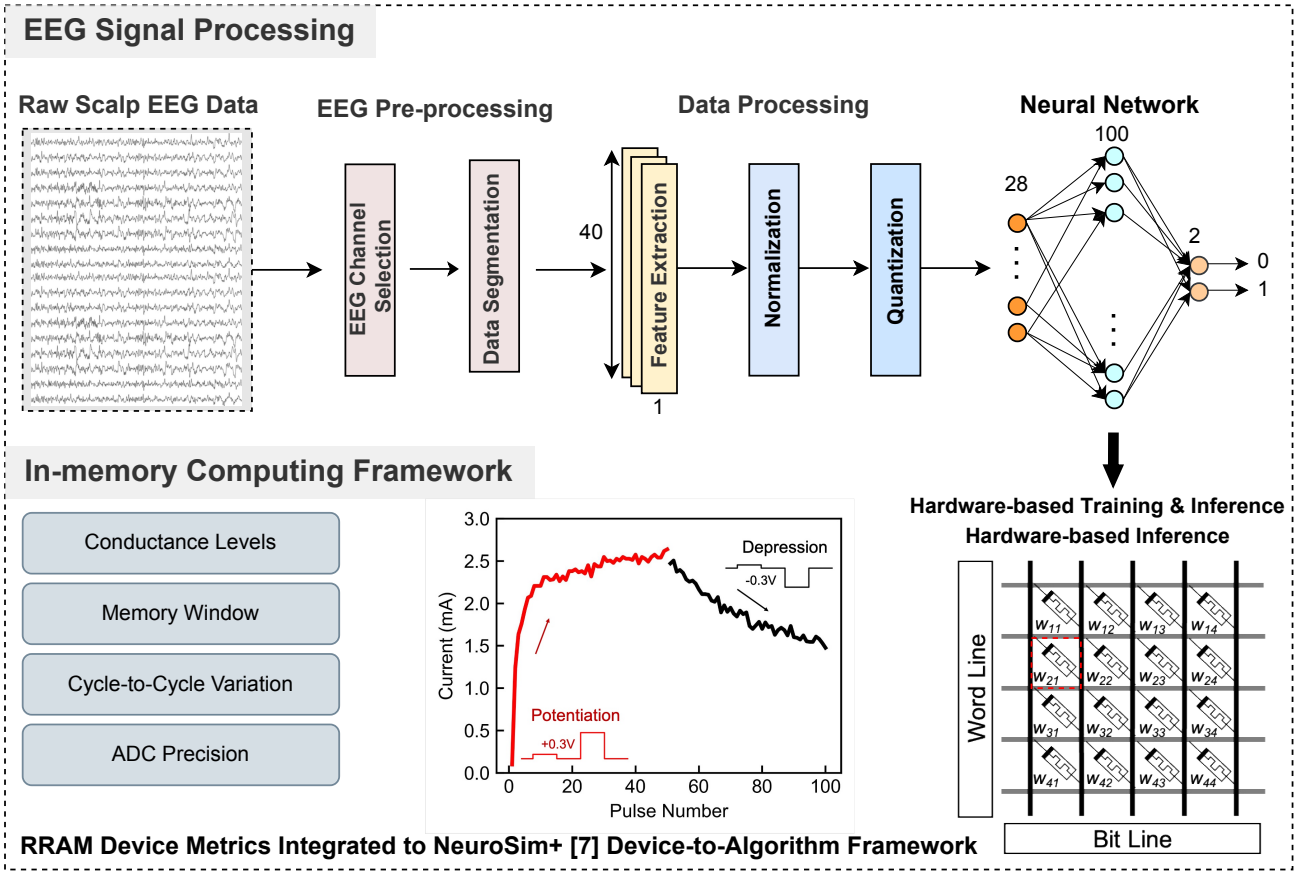


Fig. 1: Overview of the EEG Signal Processing and In-Memory Computing Framework for Benchmarking Study

B. Feature Extraction

After pre-processing the EEG data, we extracted 40 features in the time and frequency domain from each 5-second segment. Our feature extraction process is based on our prior work [8]. The time-domain features include the energy distribution, peak-to-peak values, and the number of zero-crossings. The total number of extracted features in the time-domain is 3×4, where 4 is the number of channels. The features in the frequency-domain include the spectral power intensity, which is extracted from the EEG data for 7 frequency bands from 0.1 Hz to 100 Hz using PyEEG library [9], including 0.1-4, 4-8, 8-13, 13-30, 30-50, 50-80, and 80-100 Hz.

C. Normalization and Quantization

In order to adapt the EEG data for the 1-bit input format required by the single clock-cycle process of the NeuroSim+ [7] version 3.0 (V3.0) framework, we first normalized the extracted features for each segment to a range between 0 and 1. Subsequently, these normalized data inputs were quantized to either 0 or 1, using a threshold of 0.5. This quantization process resulted in several zero values for the features of both preictal and interictal samples, particularly for lower-valued features such as mean, skewness, and kurtosis. To ensure the relevance and utility of our feature set, we subsequently removed all features that consistently registered as zero in both preictal and interictal samples. This modification allowed us to have consistent results in both the software and hardware-based training and inference configurations.

D. Multi-layer Perceptron Simulator Framework

We used the NeuroSim+ V3.0 [7] platform to simulate the effect of the RRAM device performance in predicting seizures using EEG data. Neurosim+ V3.0 is a 2-layer multi-layer perception (MLP) simulator, integrating device, circuit, and algorithm level architectures in order to emulate the software and hardware-based learning and classification tasks using the MNIST handwritten dataset [10]. As mentioned in *Section II*, we modified the CHB-MIT dataset to be represented in 1-bit. This enabled us to process the EEG data in one clock cycle using the NeuroSim V3.0 framework for our benchmarking task. Our implementation employed an FCNN with 3 layers: 40 input neurons, 28 hidden neurons, and 2 output neurons. This input to hidden neural network, 40x28, is encoded onto the RRAM crossbar array as described in [7]. The resulting vector matrix multiplication (VMM) output current is then converted to voltage and passed through the hidden to output

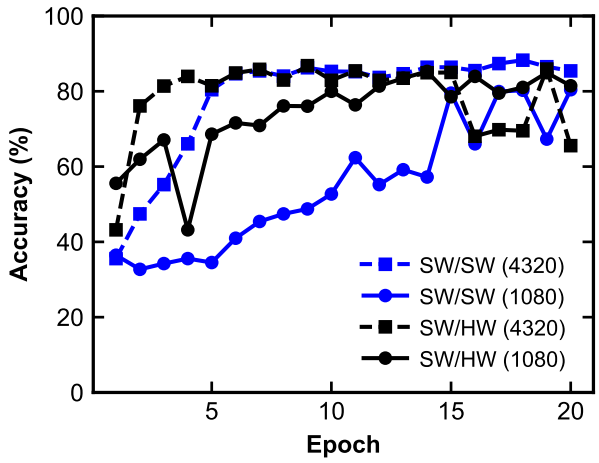


Fig. 2: Accuracy vs Epoch for SW/SW and SW/HW Configurations with Batch Sizes of 4320 & 1080

neural network of 28x2. Sigmoid was used as an activation function in both layers, and stochastic gradient descent (SGD) was used as an optimizer with a learning rate of 0.01. Our training dataset consists of 4320 EEG samples, while the test dataset includes 1440 samples. In order to evaluate the impact of critical device (conductance level, memory window) and peripheral circuitry (analog-to-digital conversion precision) metrics, we modify the default Ag-Si device configuration [12] encoded in the simulator framework, with one-transistor one-resistor (1T1R) configuration. We implement a linear weight update scheme to study the device output parameter space in isolation, without introducing additional non-idealities, except in Figure 4 where we report the effect of cycle-to-cycle (C2C) variation.

III. Results

A. Software and Hardware-Based Learning Configurations

We have evaluated three learning configurations supported by the NeuroSim+ framework i) *Fully software based training and inference (SW/SW):* Learning and inference operations are based on software without considering RRAM hardware parameters. This configuration, also referred to as software-only mode, represents the ideal setting where inherent RRAM limitations would not impact the algorithm accuracy. ii) *Software based training and hardware-based inference (SW/HW):* In this mode, the network is pre-trained using the software, and the simulator only emulates hardware during the classification/inference phase. Here, the hardware-based process involves only the feed-forward (FF) operation during inference, and RRAM conductance values, analogous to synaptic weights, are updated only once as there is no hardware-based training process involved. iii) *Hardware-based training and inference (HW/HW)*: In this configuration, both training and inference steps are implemented on hardware. The process employs both FF and backpropagation (BP) operations. During FF, input data traverse from the input layer to the output layer through VMM operations and activation functions, generating a prediction. Errors in these predictions are then adjusted through BP, where they are used to modify the RRAM conductance values using an SGD method. The simulator drives the algorithm training by emulating the hardware parameters of 1T1R RRAM crossbar matrix and the peripheral circuitry, using a subset of data batch randomly selected from the training dataset at each epoch.

Figure 2 presents the accuracy of the epilepsy prediction classification in the single patient case (Patient 9 from the CHB-MIT dataset) for SW/SW and SW/HW configurations as a function of the size of the EEG dataset included in each training epoch, batch size. We used the best-performing device parameters for the SW/HW configuration (Memory Window: 1250, Conductance Levels: 2048). For both cases, increasing the batch size increased the training rate. However, SW/SW runs exhibited a slower convergence to higher accuracy values compared to SW/HW runs. This interesting outcome is attributed to the hardware limitations of the SW/HW configuration. More specifically, the uncertainty introduced by the limitations of the RRAM device metrics can potentially help the training algorithm escape local minima, thereby enhancing overall training efficiency.

After evaluating the effect of batch size and the comparison of SW/SW and SW/HW configurations, we investigated the effect of the precision of Analog-to-Digital (ADC) conversion, represented by the numBitPartial parameter in the simulator, in the SW/HW setting. In Figure 3, accuracy values recorded at the end of the 20th epoch are reported for two batch sizes as a function of ADC precision bit. We found that the accuracy plateaus at 50% for the minimum level of 2 bits for both cases but increases with higher bit values of 4, 6, and 8. Moreover, the results are consistent for both batch sizes, with the 1080 batch showing a steadier increase in accuracy than the 4320 batch, which has a drop in accuracy at the 20th epoch, as can also be seen in Figure 2. These findings demonstrate that higher ADC precision leads to faster accuracy convergence.

B. Memory Window, Conductance Level and Cycle-to-Cycle Variation Benchmarking

We also studied the learning performance in HW/HW configuration by varying the RRAM device performance metrics and cycle-to-cycle (C2C) variation. The accuracy results of the HW/HW mode as a function of RRAM memory window, conductance level, and C2C variation metrics are presented in Figure 4. These parameters are critical in encoding the synaptic weight analogous conductance values within RRAM hardware. The memory window specifies the range in which synaptic weights can be encoded, while conductance level indicates weight quantization effects during training. Fig. 4(a) illustrates a positive correlation between increased memory window size and higher accuracy levels. However, settings with memory windows of 12.5 and 125 showed significant oscillations, indicating that stable accuracy may not be achievable within the 20-epoch training period under these condi-

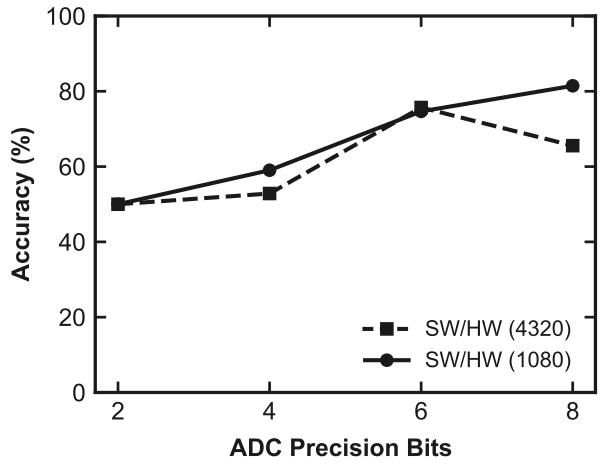


Fig. 3: Effect of ADC Precision in SW/HW Configuration for the Batch Sizes: 4320 & 1080.

tions. This instability is likely due to the extreme quantization of input and the constraints imposed by the device metrics.

Fig. 4(b) compares the results of three conductance levels, 128, 512, and 2048, representing the weight precision of 7,9, and 11 bits. Varying the weight-bit precision gradually from 2 to 12 bits reveals that achieving a stable accuracy above 80% in the studied classification task requires a minimum of 10-bit precision, given the current hardware and network parameters. Higher precision in synaptic weight encoding is facilitated by the availability of more conductance levels, allowing for fine, incremental updates as guided by the learning rate. [15] This underscores the critical role of adjusting weight-bit precision to optimize the learning algorithm. The memory window set for this experiment was as high as 12500 to decouple the impact of memory limitation from conductance level exploration, aligned with DC-based high resistance/low resistance ratio measurements reported.

Figure 4(c) demonstrates the effect of C2C variation on the weight update process, featuring a memory window of 12500 and conductance levels set at 2048. The absence of C2C variation promotes more stable weight updates, leading to higher accuracy achieved over fewer epochs. In fact, the training process is highly susceptible to C2C variations; even a minimal C2C variation of 0.5% introduces significant oscillations, and a variation of 1% compromises the consistency of the training process. These findings contrast with the previously reported image classification tasks benchmarked using RRAM hardware, where training accuracy remained above 70 % for up to 3% C2C variations [11].

C. Multiple Patients

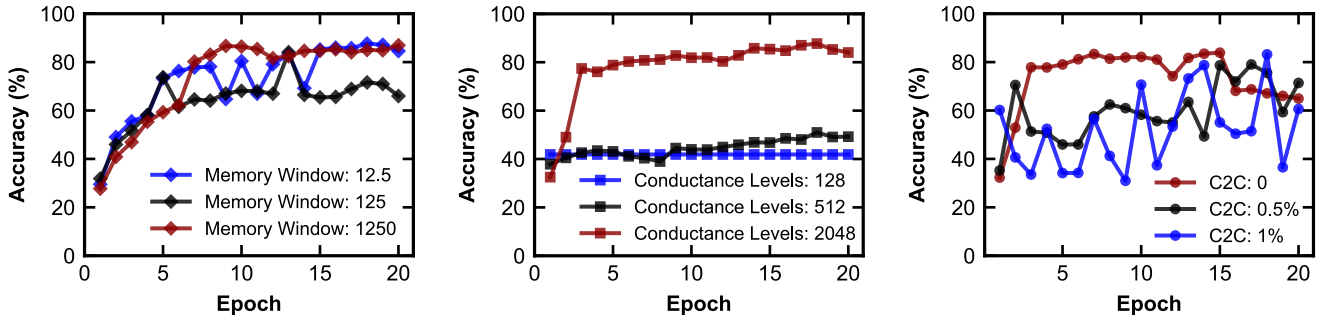
In the previous sections, we provided detailed explanations of the benchmarking process for our FFNN model using data from a single patient (Patient 9) from the CHB-MIT dataset. To assess the consistency of our model's performance, we extended our analysis to include additional patients from the dataset: Patients 1, 5, 7, 13, 22. Table I summarizes the performance of our proposed workflow, accuracy metric obtained at epoch 20, across three distinct modes: SW/SW, SW/HW, and HW/HW - for the six patients. We use two different batch sizes to accommodate the varying number of signals available from the extracted training datasets across patients. For Patients 1, 7, 13, and 22, we use a batch size of 1440, while for Patients 9 and 5, we use a larger batch size of 4320. For hardware-based learning configurations, we employ a memory window of 125 and a conductance level of 1024. These parameters were chosen to benchmark more reasonable device metrics that align with state-of-the-art values reported in the literature [15]. It can be noted that Patients 1 & 22 exhibit the highest levels of performance across the three modes, while the accuracy metrics for Patients 5, 9, and 13 remain at a lower range, with a minimum of 65.69% for HW/HW and a maximum of 85.42% for SW/SW configurations, respectively for Patient 9. Patient 7 yields the least favorable results, $\sim$50% range. As expected SW/SW and SW/HW configurations outperform the HW/HW. The limited memory window especially introduces oscillations in the accuracy per epoch for Patients 5, 7, and 9 in HW/HW mode. In repeated runs, the accuracy for these patients in HW/HW varied between 65% and 82%. Conversely, the accuracy for the other patients remained stable over the 20-epoch training period. Patients 1 & 22 consistently achieved accuracy levels exceeding 94%, whereas Patient 13's accuracy was limited to $\sim$70%.

TABLE I: Accuracy of Prediction for Multiple Patients.

Patient	SW/SW	SW/HW	HW/HW
Patient 1	97.29%	96.18%	95.83%
Patient 5	76.11%	79.13%	69.44%
Patient 7	58.54%	53.68%	50.97%
Patient 9	85.42%	65.56%	65.69%
Patient 13	70.14%	69.58%	70.00%
Patient 22	95.21%	95.56%	94.51%
Mean	**80.45%**	**76.61%**	**74.41%**

IV. Discussion

Significant variations in accuracy can potentially result from the diverse seizure types present in each patient's data. To explore the underlying data characteristics more thoroughly, we used frequency domain visualization techniques in EEGLAB for each patient. Our analysis revealed distinct patterns. For Patients 1 and 22, there were clear differences in power levels between preictal and interictal samples, allowing the model to accurately distinguish between these states without the need for complex feature adjustments or changes in electrode configurations. In contrast, such differences were not easily observed in the frequency domain for the other patients. This suggests that the data complexity is a key factor contributing to suboptimal model performance in these cases. Improving the model's accuracy for these patients may require incorporating additional features to better differentiate between preictal and interictal states, selecting alternative electrode channels, or employing more sophisticated neural network models, such as convolutional or deep neural networks.



(a) (b) (c)

Fig. 4: Impact of Device Parameters on Prediction Accuracy: (a) Memory Window, (b) Conductance Level, (c) C2C Variation

Beyond the challenges posed by data complexity and varying seizure patterns across patients, our model performance is also hindered by the inherent limitations in our benchmarking framework. These limitations include:

Quantization: Converting normalized inputs into binary values (0 or 1) to enable the single clock-cycle process oversimplifies the data, potentially leading to a loss of critical information.

Limited Features: Reduced number of input features as a result of the quantization process further limits the information needed for accurate seizure prediction.

Model Structure: The proposed FCNN architecture's limited layer number and size due to the inherent limitations of the NeuroSim+ framework may not fully capture the complexity of the data.

To improve prediction accuracy, we can adopt several strategies, such as increasing the number of input features, employing more complex neural network architectures with additional layers or neurons, to better capture the intricate patterns within the data, and exploring alternative data channels or different electrode configurations to reveal new insights. Lastly, alternative pre-processing techniques could uncover hidden patterns, especially for patients whose current performance metrics are suboptimal. By implementing these strategies, we anticipate significant improvements in prediction accuracy, which will be the focus of our future research.

V. CONCLUSION

In conclusion, our study confirms the substantial potential of the RRAM crossbar systems for predicting epilepsy seizures using EEG data. Despite the noise introduced by the strict data quantization process and the modest network topology, our framework achieves accuracy levels exceeding 80% with optimized RRAM device metrics, thanks to the meticulous domain-specific feature extraction process along with an RRAM hardware configuration of ambitious, yet feasible, high-performing device parameters.

ACKNOWLEDGMENT

G.Tutuncuoglu and M.Alhawari would like to acknowledge the following grants from NSF. 2153177: 3D Printed Application-Specific Neuromorphic Circuits: Design, Fabrication, and Implementation, for Tutuncuoglu; and 2221753: An Energy-Efficient, CMOS-based, and Scalable Mixed-Signal DNN System with Reconfigurable Crossbars, for Alhawari.

REFERENCES

[1] Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. Nat Mater 18, 309–323 (2019).
[2] Banerjee, P. N., Filippi, D. & Hauser, W. A.. "The descriptive epidemiology of epilepsy—a review." Epilepsy research 85.1 (2009): 31-45.
[3] Van Mierlo, P., et al. "Functional brain connectivity from EEG in epilepsy: Seizure prediction and epileptogenic focus localization." Progress in neurobiology 121 (2014): 19-35.
[4] Rasheed, K., et al. "Machine learning for predicting epileptic seizures using EEG signals: A review." IEEE reviews in biomedical engineering 14 (2020): 139-155.
[5] A. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, 2009
[6] Rahman, Rihat, et al. "Comprehensive analysis of EEG datasets for epileptic seizure prediction." 2021 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2021.
[7] Chen, P.-Y., Peng X., & Yu, S. "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures", IEEE International Electron Devices Meeting (IEDM), 2017, San Francisco, USA.
[8] Varnosfaderani, Shiva Maleki, et al. "A two-layer LSTM deep learning model for epileptic seizure prediction." 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS). IEEE, 2021.
[9] F. Bao, X. Liu, and C. Zhang, "PyEEG: An open source python module for EEG/MEG feature extraction," in Computational Intelligence and Neuroscience, 2011.
[10] Deng, L., The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142 (2012).
[11] Chen, P.-Y., Peng, X. & Yu, S. NeuroSim: A Circuit-Level Macro Model for Benchmarking Neuro-Inspired Architectures in Online Learning. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 37, 3067–3080 (2018).
[12] Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P., Lu, W. (2010). Nanoscale memristor device as synapse in neuromorphic systems. Nano letters, 10(4), 1297-1301.
[13] Goldberger, A., et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220." (2000).
[14] Campbell, Kristy A. "Self-directed channel memristor for high temperature operation." Microelectronics journal 59 (2017): 10-14.
[15] Rao, Mingyi, Hao Tang, Jiangbin Wu, Wenhao Song, Max Zhang, Wenbo Yin, Ye Zhuo et al. "Thousands of conductance levels in memristors integrated on CMOS." Nature 615, no. 7954 (2023): 823-829.