Seizure Detection and Prediction by Parallel Memristive Convolutional Neural Networks

Chenqi Li†, *Student Member, IEEE,* Corey Lammie†, *Student Member, IEEE,* Xuening Dong, *Student Member, IEEE,* Amirali Amirsoleimani, *Member, IEEE,* Mostafa Rahimi Azghadi, *Senior Member, IEEE,* and Roman Genov, *Senior Member, IEEE*

***Abstract*—During the past two decades, epileptic seizure detection and prediction algorithms have evolved rapidly. However, despite significant performance improvements, their hardware implementation using conventional technologies, such as Complementary Metal–Oxide–Semiconductor (CMOS), in power and area-constrained settings remains a challenging task; especially when many recording channels are used. In this paper, we propose a novel low-latency parallel Convolutional Neural Network (CNN) architecture that has between 2-2,800x fewer network parameters compared to State-Of-The-Art (SOTA) CNN architectures and achieves 5-fold cross validation accuracy of 99.84% for epileptic seizure detection, and 99.01% and 97.54% for epileptic seizure prediction, when evaluated using the University of Bonn Electroencephalogram (EEG), CHB-MIT and SWEC-ETHZ seizure datasets, respectively. We subsequently implement our network onto analog crossbar arrays comprising Resistive Random-Access Memory (RRAM) devices, and provide a comprehensive benchmark by simulating, laying out, and determining hardware requirements of the CNN component of our system. To the best of our knowledge, we are the first to parallelize the execution of convolution layer kernels on separate analog crossbars to enable 2 orders of magnitude reduction in latency compared to SOTA hybrid Memristive-CMOS Deep Learning (DL) accelerators. Furthermore, we investigate the effects of non-idealities on our system and investigate Quantization Aware Training (QAT) to mitigate the performance degradation due to low Analog-to-Digital Converter (ADC)/Digital-to-Analog Converter (DAC) resolution. Finally, we propose a stuck weight offsetting methodology to mitigate performance degradation due to stuck R_{ON}/R_{OFF} memristor weights, recovering up to 32% accuracy, without requiring retraining. The CNN component of our platform is estimated to consume approximately 2.791W of power while occupying an area of 31.255mm2 in a 22nm FDSOI CMOS process.**

***Index Terms*—CNN, Seizure Detection, Seizure Prediction, EEG, RRAM, Memristive Crossbar Array**

I. Introduction

EPILEPSY is a common neurological disorder that affects approximately 1% of the world's population [1]. A seizure is characterized by excessive firing of neurons in the brain, while epilepsy is a medical condition that involves recurrent seizures [2]. As the underlying occurrence mechanism of epilepsy is not well understood [3]–[5], it requires experimental methods of treatment that rely on accurate detection and prediction systems, as depicted in Fig. 1.

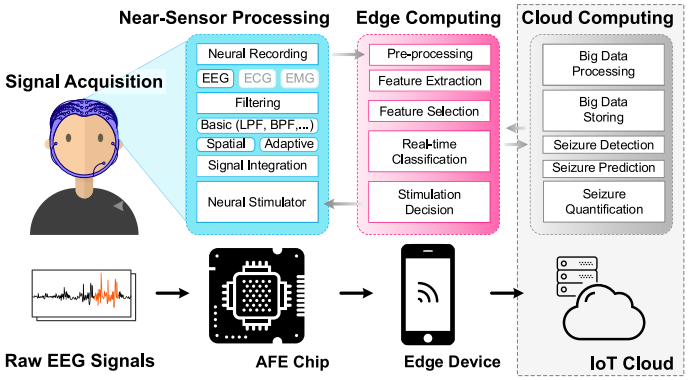


Fig. 1. An overview of a typical epileptic seizure detection and prediction system. Acquired EEG signals are sampled and processed near-sensor using an Analog Front End (AFE), prior to being sent wirelessly to edge device(s) for real-time pre-processing and feature extraction. Features can then be fed into ML and/or DL architectures, residing either on the IoT edge or in the IoT cloud, which perform epileptic seizure detection and prediction.

EEG is the most common method used to monitor the electrical activities of the brain, and can be used to detect and predict seizures. There have been numerous applications of traditional ML algorithms, such as Support Vector Machines (SVMs), k-Nearest Neighbor (kNN), and Random Forest (RF) classifiers to classify ictal (seizure), preictal (prior to a seizure) and non-ictal (non-seizure) signals using EEG recordings. Despite being able to achieve high accuracies, these approaches require the manual extraction and selection of features in the time- or frequency-domain [6]. The optimal choice of such feature extractions are largely unknown, experimental, and dependant on specific patient signatures, such that there is no one-fit-all solution.

Compared to traditional seizure classification algorithms, DL-based algorithms have more advantages in complex EEG signal feature extraction, as they do not require feature engineering, and are capable of outperforming traditional ML algorithms for epileptic seizure detection and prediction tasks [7]. However, when these DL systems are implemented using CMOS, there are problems such as large scale, high calculation energy consumption and high delay, which hinder

†**These authors contributed equally.**
Corresponding authors: M. Rahimi Azghadi and A. Amirsoleimani.
Chenqi Li, Xuening Dong, and Roman Genov are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada. email: chenqi.li@mail.utoronto.ca, xuening.dong@mail.utoronto.ca, roman@eecg.utoronto.ca
Corey Lammie and M. Rahimi Azghadi are with the College of Science and Engineering, James Cook University, QLD 4811, Australia. e-mail: corey.lammie@jcu.edu.au, mostafa.rahimiazghadi@jcu.edu.au
Amirali Amirsoleimani is with the Department of Electrical Engineering and Computer Science, York University, Toronto, Canada. e-mail: amirsol@yorku.ca

their efficacy; especially in resource-constrained environments.

In order to solve this kind of problem, this paper proposes a neuromorphic calculation strategy based on a novel In-Memory Computing (IMC) RRAM architecture, which utilizes analog crossbars. Computer designers have traditionally separated the role of storage and compute units. The IMC paradigm blurs this distinction, and imposes the dual responsibility on memory substrates: storing and computing on data for massively parallel computing [8]. By exploiting the physical characteristics of emerging analog device technologies, analog crossbars can be used to perform Vector-Matrix Multiplications (VMMs), the most dominant operation in CNNs, in as little as $\mathcal{O}(1)$ [9], [10], significantly reducing the computational complexity during inference operations. Our specific contributions are as follows:

1) To the best of our knowledge, we are the first to parallelize the execution of convolution layer kernels on separate analog crossbars to address the computational bottleneck of CNNs, enabling 2 orders of magnitude reduction in latency compared to current SOTA hybrid Memristive-CMOS DL accelerators;
2) We reduce the number of required parameters by 2-1,600x and 5-2,800x for epileptic seizure detection and prediction tasks using deep learning models, while still achieving SOTA performance;
3) We provide a comprehensive benchmark for hardware memristor-based seizure prediction/detection systems by simulating, laying out, and determining hardware requirements of the CNN component of our system;
4) We propose a simplified stuck weight offsetting methodology for mitigating severe degradation of system performance due to stuck $R_{\text{ON}}/R_{\text{OFF}}$ memristor weights. We demonstrate that our method is capable of achieving up to 32% performance recovery, without requiring retraining, while incurring minimal hardware and computational overhead.

To promote reproducible research, all of our simulation codes are made publicly accessible[1]. The rest of the paper is structured as follows: In Section II, we overview and discuss related work. In Section III, we present our epileptic seizure detection and prediction system. In Section IV, we overview and discuss our software methodology. In Section V, we overview and discuss our hardware simulation methodology. In Section VI, we present and discuss our results. Finally, we conclude the paper in Section VII.

II. Related Work

In this section, we present an overview of related work using parallel CNNs and related work using traditional and neuromorphic ML architectures for the detection and prediction of epileptic seizures using EEG and Intracranial Electroencephalography (iEEG) signals.

A. Parallel CNNs

Parallel CNNs are composed of one or many convolutional layers, which are executed in parallel and have been previously used in many applications. For example, in the ResNeXt [12] family of architectures, parallel blocks containing convolutional layers were used to increase network width, which can decrease the time required to train a CNN [13]. When performing multi-modal DL, parallel convolutional layers can be used to process different inputs in parallel [14], in order to improve network throughput. Specifically for epileptic seizure detection and prediction tasks, parallel convolutional layers have been used to learn high-level representations simultaneously [15]. By parallelizing convolutional operations, inference time is greatly reduced compared to current SOTA architecture that rely on sequential convolution layers, as convolution layers form the bottleneck of CNN inference.

B. Traditional EEG-based Seizure Detection and Prediction Algorithms

As early as 1996, initial attempts were made to detect seizures using EEG signals and traditional ML approaches. Using a combination of Artifical Neural Networks (ANNs) and wavelet transforms, sensitivity values of 76% [16] and 97% [17], were reported using standardized datasets. In the late 2000s and early 2010s, SVMs encountered growing interest. Namely, when using SVMs in combination with feature extraction methods such as high-order spectra analysis, wavelet transforms, Fast Fourier Transforms (FFTs), wavelet decomposition and least-squares parameter estimators [18]–[27], promising sensitivity, specificity, and accuracy values $\geq$98.5% were achieved. More recently, advances in the DL domain using CNNs and Recurrent Neural Networks (RNNs), have further benefited seizure detection algorithms. Current SOTA models are capable of achieving accuracy ranging from 95-100% [28]–[31] across multiple datasets.

Early efforts for seizure prediction started in 1970s, where seizure warning systems were designed with logic circuitry to classify extracted features from a series of filters and analog circuitry [32], [33]. To varying degrees of success, a variety of methods have been proposed, including a rule-based method using univariate measures [34], spike rate analysis [35], positive zero-crossing intervals analysis [36], statistical dispersion measures [37], multidimensional probability evolution [38], circadian concepts via probabilistic forecasting [39], and a combination of reinforcement learning, online monitoring and adaptive control theory [40]. Similarly to seizure detection, many DL techniques have also been applied. Notable contributions include the combination of CNNs and RNNs, capable of achieving 99.6% accuracy and a False Positive Rate (FPR) of 0.004 per hour [41]. Moreover, supervised deep convolutional autoencoder and bidirectional long short-term memory networks have been used to achieve accuracy, sensitivity, specificity, and precision values between 98-99%, with F1-values $\geq$0.98. More recently, augmented DL network architectures have been used to reduce computational complexity for operation in resource-constrained environments. One such approach, which employs CNNs with minimizing channels, is capable of achieving 99.47% accuracy, 97.83% sensitivity, 92.36% specificity, with a FPR of 0.0764 [42]. Finally, Siamese models have been used to achieve 88-91%

[1]https://anonymous.4open.science/r/7f3fd487-2e87-4d47-8d26/

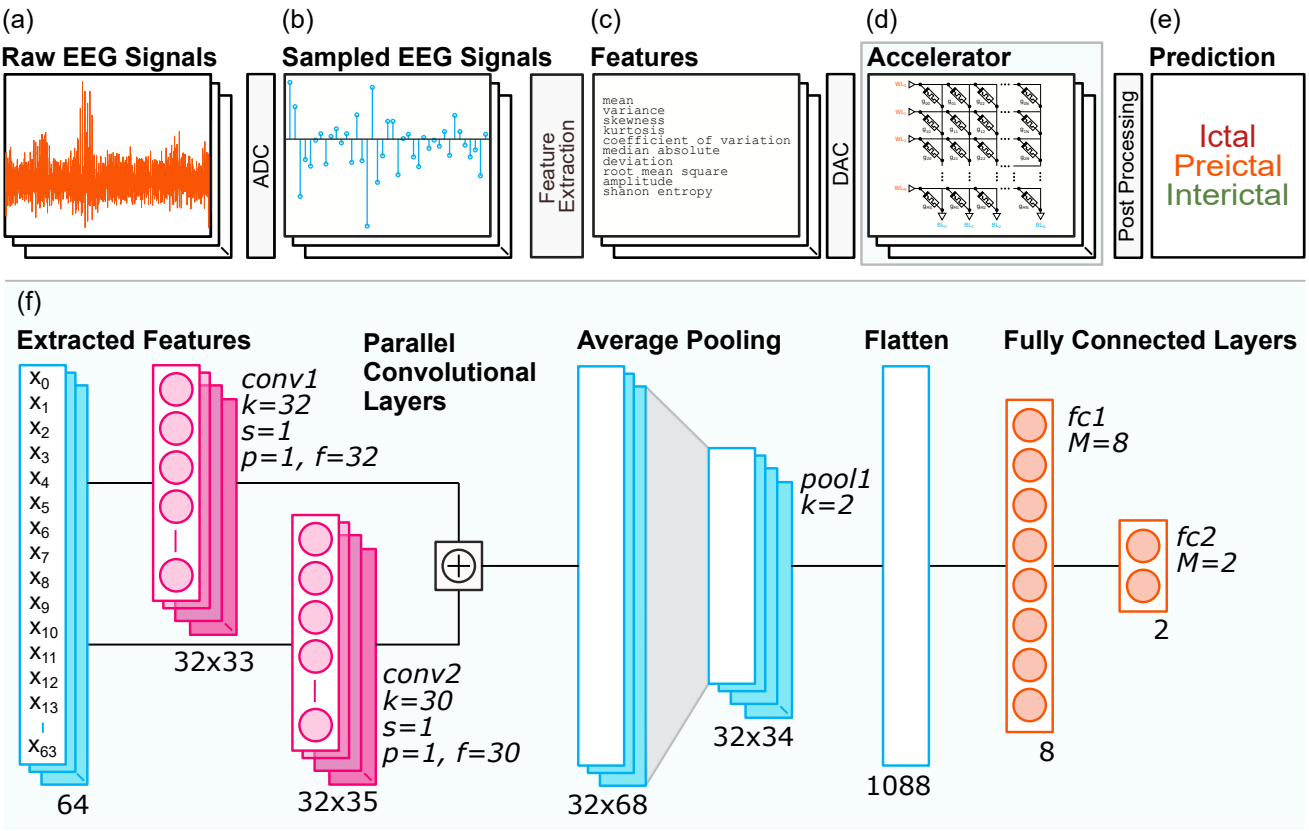


Fig. 2. A high-level system architecture overview. (a) Raw EEG signals are sampled and digitized using ADCs. (b) Features are extracted from sampled EEG signals. (c) Extracted features are fed into a memristive DL accelerator. (d) Accelerator outputs are processed. Fig. 3 depicts the detailed hardware implementation of the accelerator. (e) Processed accelerator outputs are used to determine interictal, preictal, and ictal states. (f) The novel neural network architecture used consists of two parallel 1d-convolutional layers, one average pooling layer, and two fully connected (dense) layers. N is used to denote the batch size, i.e., the number of batches presented to the network in parallel. f denotes the number of filter. k determines the filter size. s denotes the stride length. p denotes the padding. M denotes the number of output neurons for each fully connected layer. Parts of this figure are derived from [11].

accuracy on the CHB-MIT dataset [43]. We refer the reader to [44] for a comprehensive survey of EEG seizure detection and prediction algorithms.

C. Hardware Implementations of EEG-based Seizure Detection and Prediction Algorithms

Many hardware implementations of epileptic seizure detection and prediction algorithms have been reported using a variety of technologies; namely Field Programmable Gate Array (FPGA), CMOS and Very-large-scale Integration (VLSI) [45], [46]. Complementing traditional hardware implementations, IMC architectures, which use memristive crossbar arrays to perform repetitive operations in-memory, have gained increasing popularity in recent years. Kudithipudi *et al.* implemented a neuromemristive reservoir computing architecture to achieve 90% accuracy and Merkel *et al.* achieved 85% accuracy [47], [48]. Nature-inspired memristive Cellular Automata (CA) was implemented by Karamani *et al.* to emulate epilepsy-related phenomena in the brain [49].

Recent works by Liu *et al.* implemented Finite Impulse Response (FIR) filter bank on memristive crossbar array to achieve 93.46% seizure detection accuracy and obtained 95% accuracy by using a memristive crossbar based signal-processing stage combined with linear discriminant classifier [50]. Lammie *et al.* pioneered the implementation of CNNs for seizure prediction using memristor arrays, achieving 77.4% sensitivity and 0.85 Area Under the Receiver Operating Characteristic Curve (AUROC) on the CHB-MIT dataset [11].

Seizure is a chronic, recurring condition that can mostly be prevented through medication before onset [51], but even with the best medications, 30% of the patients are drug-resistant [52]. Closed-loop brain stimulation has been found to mitigate and even improve symptoms [53], [54], but unpredictability of seizure requires a closed-loop prediction system to provide accurate warning with adequate preparation time for stimulation [55]. This calls for the need for fast, low-latency computations, as the changes within the patients can be noticed early-on, in order to start treatments early to improve safety and quality of life [56]. In doing so, symptoms and subsequent effects can be minimized, including anxiety and social exposition [57]. The major limiting factor of seizure detection and prediction algorithms is the reliance on patient specific features, leading to undesirable results when generalized to other patients in the real world [58]. With energy efficient computations, it enables the deployment of such systems within wearable devices, so that it can be coupled with the stimulation system, as well as allowing data for a patient to be collected in the long-term to further improve model's predictions by fine tuning the model to better recognize patient-specific signatures [59].

It is known that convolutional layers are the bottlenecks

of CNNs. According to Cong et al., convolutions make up more than 90% of CNN inference [60]. Therefore, accelerating convolution is pivotal to efficient CNNs for future seizure detection/prediction systems. Note that all existing hardware implementations of CNN memristive accelerators focus on sequential CNNs. Memristive crossbar acceleration of parallelized convolution layers and blocks, found in many CNN architectures such as ResNeXt [12], are explored in this work to further reduce inference latency.

III. Seizure Detection and Prediction System

In this section, we present our seizure detection and prediction system. As shown in Fig. 2, our system comprises of five stages, depicted using Fig. 2(a)-(e). As the same network architecture, depicted in Fig. 2(f), is used for both detection and prediction, and networks are bench-marked using multiple datasets, our proposed system can be reconfigured for both epileptic seizure and prediction tasks. While we briefly detail and discuss signal acquisition and pre- and post-processing stages, here-on-in, the scope of this paper will be largely confined to the accelerator step described in Fig. 2(d). We leave a detailed hardware description and evaluation of other stages to future work.

A. Parallel Convolutional Neural Network Architecture

The primary constraint put on our design was a fixed modular tile size of 64×64. Practically, passive memristor-based analog crossbar tiles of sizes up to 128×64 have been used to perform VMMs [9], however such designs have only been demonstrated using pseudo-crossbars having micron-size electrodes. Such limitations in the maximum viable size are a serious computational scalability challenge with electrodes in the tenth of nanometer range that would prevent sinking large currents through them [61]. Recently, a 4K memristor analog-grade passive crossbar circuit has been fabricated [62], which comprises several modular 64 x 64 passive crossbar tiles with 99% functional nonvolatile metal-oxide memristors. From an original exploratory investigation, it was determined that for the RRAM device being modelled, the largest feasible modular tile size which is able to be programmed using a write-verify scheme was 64×64. Consequently, this fixed modular tile size was used in our designs to minimize the power and area overhead of peripheral circuits and tile interconnects, which are much larger when smaller fixed modular tiles are used.

B. Model Search and Selection

Most current state-of-the-art CNNs employ sequential convolution layers, whereby subsequent convolution operations are dependent on results from previous layers. However, in parallel CNNs, convolution layers can be processed simultaneously, enabling the use of multiple crossbars at the same time. In addition, parallel convolution layers with different kernel sizes enable the network to extract features of varying receptive fields, providing the fully connected layers a diverse and yet compact representation of the features for classification; enabling a reduction in network parameters required.

Algorithm 1 Model Search and Selection Methodology

Input: Fixed modular crossbar tile size ($m \times n$), OBJ_{max}, objectives to minimize, OBJ_{min}, additional hardware design constraints, $\mathbf{w}$.

Output: Optimized network architecture $(L, D, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where L is the number of convolutional layer blocks, D is the number of fully connected layers, $\boldsymbol{\alpha}$ is a vector containing the sizes of the first kernel for each convolutional layer when parallel convolutional layer execution is performed, and $\boldsymbol{\beta}$ is a vector containing the number of output neurons for each fully connected layer

minimize $\text{OBJ}(m, n, L, D, \boldsymbol{\alpha}, \boldsymbol{\beta})$ subject to $\mathbf{w}$.

procedure NETWORK_ARCHITECTURE($m, n, L, D, \boldsymbol{\alpha}, \boldsymbol{\beta}$)
  for $l = 0$ to $L - 1$ **do** ▷ For each convolutional layer
    $\mathbf{C_{in}}_l = m$ ▷ Input channels
    $\mathbf{C_{out}}_l$ = floor(n / 2) ▷ Output channels
    if parallel convolutional layer execution **then**
      $\mathbf{k}_{l0} = \boldsymbol{\alpha}_l$, $\mathbf{k}_{l1} = m - 2 - \boldsymbol{\alpha}_l$ ▷ Set kernel sizes
    else
      $\mathbf{k}_l = m - 1$ ▷ Set kernel size
    end if
  end for
  for $d = 0$ to $D - 2$ **do** ▷ For each fully connected layer
    $\mathbf{m}_d = \boldsymbol{\beta}_l$ ▷ Set number of output neurons
  end for
  $\mathbf{m}_{D-1} = 2$ ▷ Last layer
end procedure

function OBJ($m, n, L, D, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}$)
  maximize EVAL(Net) and **minimize** PARAMS(Net), ▷ i.e., determine L, D, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, where EVAL determines the validation accuracy, and PARAMS determines the total number of network parameters
  where,
  Net = NETWORK_ARCHITECTURE($m, n, L, D, \boldsymbol{\alpha}, \boldsymbol{\beta}$)
return OBJ_{min}(Net)
end function

As shown in Fig. 2, our proposed CNN architecture consists of two parallel convolution kernels. Algorithm 1 formalizes the methodology used to search for and select the employed model. For our selected model, latency was minimized using OBJ_{min}. L, D, and $\boldsymbol{\beta}$ were fixed to values determined empirically using a preliminary exploratory analysis, and $\boldsymbol{\alpha}$ was optimized as per Algorithm 1. The following additional hardware design constraints were imposed for our design: all convolutional layers must be capable of fitting onto one modular crossbar tile, and the total number of required modular crossbar tiles must not exceed 8.

As the convolution operation bottlenecks CNN inference, the size of kernels used in parallel convolution layers need to be carefully considered to optimize both network performance and latency. In our proposed architecture, shown in Fig. 2(f), we have two parallel convolution layers and one average pooling layer, comprising one convolutional block. To parallelize the two convolution layers, it would be necessary to map the weights of the two convolution layers onto two separate cross-

bars. As a design choice, we wanted to retain the flexibility of mapping both convolution layers onto the same crossbar, if space complexity is prioritized over latency. Therefore, during the kernel size search, we imposed a constraint of 62, i.e., $m-2$, for the sum of convolution kernels, as 2 additional rows are designated for implementing the bias for both parallel convolution layers.

When denoting the kernel size of the first parallel convolutional layer as α, the kernel size of the second parallel convolutional layer can be expressed as $62-\alpha$. To determine the optimal network architecture, the University of Bonn's EEG seizure dataset [63] was used. Specifically, a 80:20 train validation split was employed, and EVAL(Net) was used to determine the 5-fold cross validation accuracy. Seed values of 32 and 8 were arbitrarily set for the network architecture search, to ensure reproducibility of results, and to reduce bias between search and validation.

Empirically, $L=1$, $D=2$, and $\boldsymbol{\beta}$=[8,] achieved substantial performance. For the single convolutional block, $\boldsymbol{\alpha}_0$ was varied between 31 and 60. A validation accuracy of 100% was achieved for all values of $\boldsymbol{\alpha}_0$, except for $\boldsymbol{\alpha}_0=60$, which achieved an optimal validation accuracy of 99.375%. This is not surprising, as the window size of input data is only 64. Therefore, convolution kernel sizes of 60 and 2 provides two extreme and dramatically different receptive fields. In particular, a kernel size of 2, which corresponds to around 10ms of data at 173.61Hz, is likely insufficient to capture local correlation and learn seizure characteristics. The final model was chosen using Occam's razor principle, whereby the simplest model is the best model. Consequently, a kernel size of 32 was selected, as a kernel size 31 would be the simplest to implement due to symmetric convolution kernel sizes; however 32 provides a more diverse receptive field. To further demonstrate the advantage of varied kernel size, a 5-fold cross-validation was performed using a) 64 filters of kernel size 31 b) two parallel convolution layers each with 32 filters of kernel size 30 and 32 (see Fig. 2). It was observed that both networks are capable of achieving accuracy varying between 99.61% to 99.83%, but varied kernel size leads to +0.03%, -0.01%, +0.02% change in performance on Bonn, SWEC-ETHZ and CHBMIT datasets, respectively, compared to using 64 filters of kernel size 31. Although a small degradation in performance is observed for SWEC-ETHZ dataset, improvements are observed for both Bonn and CHBMIT dataset. A net improvement is observed for both seizure detection and prediction using a varied kernel size, while both experiments employ an identical number of weights.

C. Hardware Architecture Hierarchy

In Fig 3, we present our hardware architecture hierarchy. The processing engines comprises 7 memristive crossbar array tiles, as well as I/O registers, eDRAM buffers, and peripheral circuits for ReLU, subtract, and average pooling. We present two configurations for our tile, Time-Division Multiplexing (TDM), and parallelized. In the TDM case, each tile contained a S+H and an ADC for reading out column currents, and one

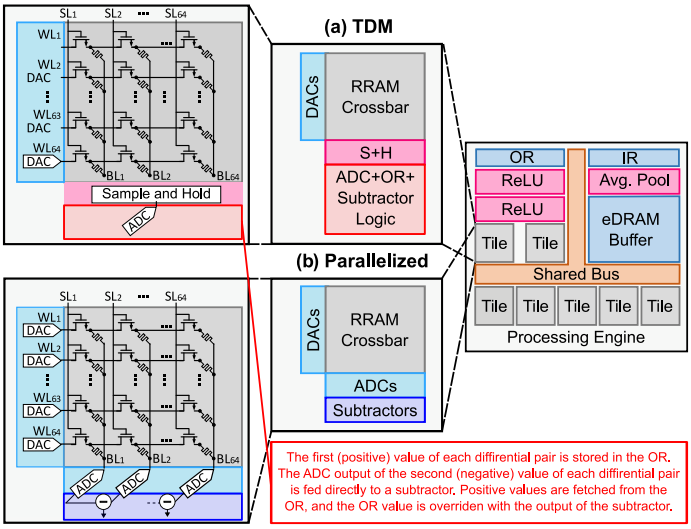


Fig. 3. Architecture hierarchy of our memristive DL accelerator with (a) TDM and (b) Parallelized Implementation.

DAC per row for reading inputs in parallel, as shown in Fig. 3(a). In the parallelized case, each tile contains 64 ADCs, as shown in Fig. 3(b).

IV. Software Methodology

To train and evaluate our epileptic seizure detection and prediction system, we benchmarked our system using one epileptic seizure detection task and two epileptic seizure prediction tasks. For epileptic seizure detection, the University of Bonn's EEG seizure dataset [63] was used. For epileptic seizure prediction, the CHB-MIT Scalp EEG [64], and the long-term SWEC-ETHZ iEEG [65] datasets were used.

To perform epileptic seizure detection and prediction, EEG and iEEG samples can be categorized as either ictal, interictal or preictal. Ictal samples indicate the presence of a seizure, interictal samples are periods between seizures, and preictal samples can be used to detect the onset of a seizure. For epileptic seizure detection, binary classification is performed between ictal and interictal samples. For epileptic seizure prediction, binary classification is performed between preictal and interictal samples. For both epileptic seizure detection and prediction tasks, on account of unbalanced classes, 5-fold cross validation was used to train and validate our network architecture.

A. Training and Evaluation Methodologies

1) Epileptic Seizure Detection: The University of Bonn's EEG seizure dataset is comprised of 5 sets (A-E), where set A is normal with open eyes, set B is normal with closed eyes, set C and D is seizure free intervals, and set E is seizure only activities. Each set contains 100 single-channel EEG time series of 23.6 seconds, with 4,096 samples in each time series. All data were collected at 173.61 Hz, at a resolution of 12 bits. To perform binary classification between ictal and interictal samples, all samples from sets A and E were used.

TABLE I
OVERVIEW OF CASES USED TO PERFORM EPILEPTIC SEIZURE PREDICTION FROM THE CHB-MIT SCALP EEG (CHB-MIT) AND THE LONG-TERM SWEC-ETHZ IEEG (SWEC-ETHZ) DATASETS.

Patient	Seizures	Interictal Hrs.*	Preictal Hrs.*	Interical Smp.†	Preictal Smp.◇	Synthetic Preictal Smp.◇
CHB-MIT						
1	7	33.74	0.43	1,898	24	42
2	3	32.85	0.14	1,848	8	14
3	7	30.86	0.39	1,736	22	37
5	5	33.85	0.30	1,904	17	30
8	5	14.93	0.36	840	20	3
SWEC-ETHZ						
1	2	19.91	1.00	1,120	56	108
2	2	19.91	1.00	1,129	56	108
3	4	29.87	1.99	1,680	112	216
5	4	29.87	1.99	1,680	112	216
6	8	69.69	3.48	3,920	196	430

*Hours. †Samples.

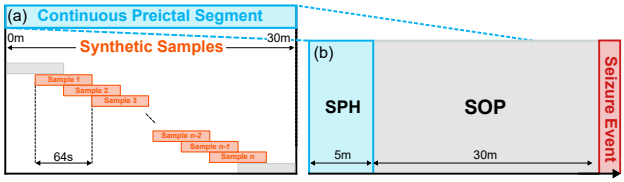


Fig. 4. Depiction of (a) our adopted overlapped sampling technique extracting n samples from a continuous preictal segment, and (b) the SPH and SOP terms. As can be seen, continuous preictal segments are extracted during the SPH. All preictal samples that occur during the SOP period are discarded.

Both sets (A and E) were divided into samples of 64 seconds periods and randomly shuffled. No augmentation and pre-processing techniques, such as normalization, were performed, as CNNs are capable of automatic feature extraction from time-series data and are robust to noise. The lack of need for pre-processing steps implies reduced hardware complexity to perform such operations. Using the network model (with optimal kernel sizes determined in Section III-B), a 5-fold cross-validation strategy was used to determine network's performance. To determine performance, the mean of left out set accuracy, sensitivity, specificity, false-positive rate and the AUROC across folds of 5-fold cross-validation were reported.

2) Epileptic Seizure Prediction: The CHB-MIT Scalp EEG, and the long-term SWEC-ETHZ iEEG datasets were used. The CHB-MIT Scalp EEG dataset comprises of 23 cases, which were collected from 22 subjects (5 males, ages 3–22; and 17 females, ages 1.5–19). The last case was obtained 1.5 years after the first, from one of the female subjects [64]. All signals were sampled at 256Hz with 16-bit resolution, using 23-26 electrodes. During data acquisition, no augmentation steps were performed.

The long-term SWEC-ETHZ iEEG dataset comprises of 18 patients with pharmaco-resistant epilepsy, who were evaluated for surgery at the Sleep-Wake-Epilepsy-Center (SWEC) of the University Department of Neurology at the Inselspital Bern [65]. All signals were sampled at either 512Hz or 1025Hz with 16-bit resolution, using 26-100 electrodes. During data acquisition, after analog-to-digital conversion, a digital band-pass filter was used to filter signals between 0.5 and 150Hz using a fourth-order Butterworth filter. Moreover, forward and backward filtering was applied to minimize phase distortion.

Due to computation burden of crossbar simulation, we report the performance using the first 5 viable cases of the the CHB-MIT Scalp EEG and long-term SWEC-ETHZ iEEG datasets, reducing the computation required, similar to [15], [66]. In Table I, we present an overview of all cases used to perform binary classification between preictal and interictal samples. A case was categorized as viable if it contained valid labels (namely time-stamps) and data files (i.e., no recording files were missing or corrupt). For both datasets, the first 22 channels of each patient were extracted and used. All signals were down-sampled to 256Hz, and a window size (batch size) of 64s was used when extracting samples. After discarding seizures that occur in the first 20-minute monitoring period, a Seizure Occurance Period (SOP) of 30m and a Seizure Prediction Horizon (SPH) of 5m were used to extract and label preictal samples for all cases; both of which have previously demonstrated significant performance [66]. These terms are defined visually in Fig. 4. Interictal samples were extracted from one hour recording segments containing no seizures (ictal samples) to reduce class inbalance during training.

Next, 176 features per sample were extracted (8 per channel per window/batch interval): the mean, variance, skewness, kurtosis, coefficient of variation, median absolute deviation of EEG amplitude and Root Mean Square Amplitude (RMSA), and the shannon entropy. Since the input size of the proposed network is 64, the dimensionality of the input data needed to be reduced. A correlation analysis was first performed across the 176 extracted features, but no particular channel could be removed as no strongly correlated channels were discovered. Using Principal Component Analysis (PCA), linear dimensionality reduction via Singular Value Decomposition (SVD) enabled the projection of data to lower dimensional space of 64 principal axes. During training, synthetic preictal samples were generated using an overlapped sampling technique inspired by [44], by sliding a 64s window with a stride of 32s across continuous preictal segments extracted during the SPH period, as depicted in Fig. 4. The same cross-validation training and evaluation strategy and metrics as described in Section IV-A1 was employed.

V. HARDWARE METHODOLOGY

In this section, we discuss our device technology selection, memristor crossbar array implementations of CNNs, and present our adopted hardware simulation methodology.

A. Device Technology Selection

Computing with charge-based computing devices is attractive due to their technological maturity, even though they have a relatively large area footprint even at advanced technology nodes and face severe scaling challenges [67]. Resistance-based memory, in contrast, can be scaled to the nanometer scale, and has the potential of forming cross-point structures without using access devices, achieving ultra high density. RRAM devices are used in our design, as they are widely considered to be the most promising emerging resistance-based memory technology- they operate faster than Phase-Change Memory (PCM), have a simpler and smaller cell structure than Magnetoresistive Random-Access Memory (MRAM) and Conductive Bridging Random-Access Memory (CBRAM)

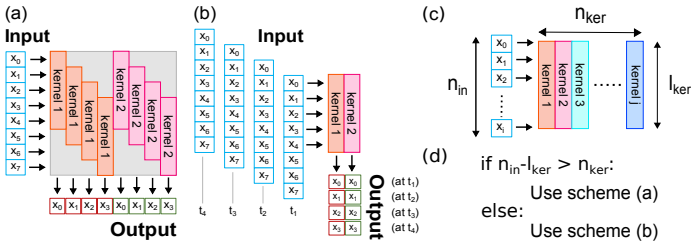


Fig. 5. A comparison of possible mapping schemes. (a) visualizes the staggering mapping of convolution weights, which is commonly adopted due to its ability to produce all results within a single pass through the crossbar array. (b) visualizes our proposed mapping scheme, without staggering of convolution weights and sparsity in crossbar, at the cost of increased read/write operations. (c) provides a comparison of methods (a) and (b), visualizing when one method should be chosen over the other.

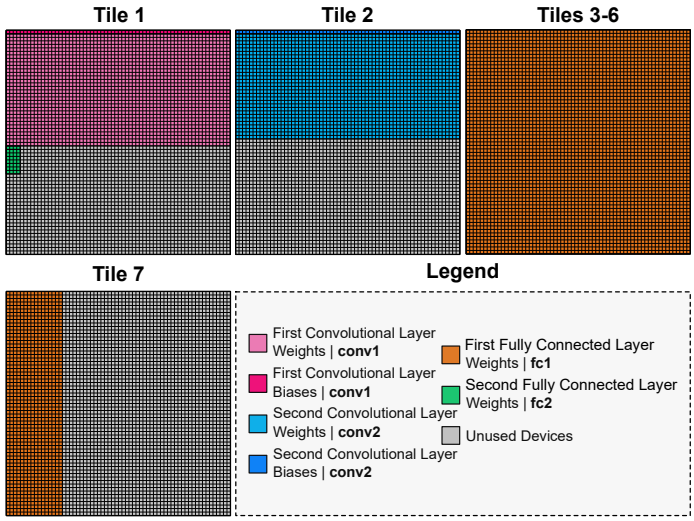


Fig. 6. The crossbar parameter mapping layout adopted. Seven 64 × 64 modular crossbar tiles are utilized. Bias terms of fully connected layers, and the single pooling layer, `pool1`, are computed using additional digital circuitry. To reduce the number of unused devices, parameters of different layers are shared between tiles.

devices, and are made of materials that are common in semiconductor manufacturing [67].

B. Memristor Crossbar Array Implementations of Parallel CNNs

Consider the conductance values of a crossbar array as a matrix and input voltages to a crossbar as a vector. The output current from the crossbar, determined using Kirchoff's and Ohm's Law represents the result of the VMM. Such operations form the core of CNNs. Being able to accelerate and parallelize them would facilitate the real-time operation of deeper and heavier neural networks for epileptic seziure detection and prediction in resource-constrained hardware [68].

To represent signed weight matrices on memristive crossbar arrays, as negative conductance values cannot be expressed using analog memristive devices, a differential mapping scheme was adopted, where two columns of memristors are chosen to represent positive and negative weights, respectively. The signed output is thus the arithmetic difference of current from both columns. In the case of 1D CNNs, fully connected and convolutional layers can be decomposed into a series of dot products between inputs, represented as voltages, and weights, represented as memristive conductance. For convolutional layers, the `im2col` algorithm [69] can be used to map convolutional kernels onto separate crossbar columns. With a single pass, m 1D convolutions can be performed simultaneously, where m represents the number of columns. Average pooling and ReLU operations are performed using additional digital circuitry.

C. Hardware Simulation Methodology

Based on existing literature from Section II-C, all mapping of convolution kernels onto crossbars are sparse, whereby the convolution kernels form a sparse diagonal matrix, as depicted in Fig. 5(a). This naive approach is extremely space demanding, as the kernels are staggered multiple times throughout the crossbar array, rendering a lot of memristive cells unused. To reduce the space requirement of mapping scheme (a), one possible approach is to build upon the input-stationary concept. One may remap the crossbar weights during inference and replace them with different kernel weights, while reusing the input fetched from memory.

On the other hand, one may build upon the weight-stationary concept, as depicted in Fig. 5(b). In this scheme, convolution kernels can be mapped without staggering before inference. For kernels to convolve against different parts of the signal, the input signal slides. The bottleneck of this approach now lies within fetching input data, requiring additional read/write operations on the peripheral of the crossbar compared to mapping scheme (a). The weight-stationary approach is more efficient compared to the input-stationary approach, as crossbar weight writes can be very time and energy consuming, compared to fetching of inputs and staggering them with shifting circuitry. Fig. 5(c) provides visualization of when one scheme should be adopted over the other.

A comparison of the naive approach and our proposed weight-stationary approach is performed for our network architecture in Table II. As can be observed, the number of memristor cells required for scheme (b) (depicted in Fig. 5 (b)) is significantly smaller, due to the compact nature of the mapping. This comes, however, at the cost of 33x increase in computation. When taking sparsity, i.e. unused memristors depicted by the gray background in Fig. 5 (a), into consideration, scheme (b) demonstrates even more significant reduction, i.e. 63x-73x fewer memristors required, while the computation increase remains constant. Unlike convolutional layers, fully connected layers do not involve sliding of signals, so VMMs for fully connected layers were implemented using the naive scheme (a). Using scheme (b), we mapped convolutional kernels within our trained network onto crossbars tiles of 64×64. While scheme (b) was chosen for our hardware design, if scheme (a) were chosen with different n_{ker} and l_{ker} values, or the added space complexity is not of concern, the staggered weights of scheme (a) would enable all rows of the crossbars to be employed simultaneously. By choosing the input size of our

TABLE II
CROSSBAR MAPPING COMPARISON FOR SPACE AND COMPUTATION TRADE-OFF USING SCHEMES (A) AND (B) IN FIG. 5.

Layer	Number of Memristor Cell Required				Number of Memristor Cell Required Inc. Sparsity			
	Scheme (a)	Scheme (b)	Area Reduction	Computation Increase	Scheme (a)	Scheme (b)	Area Reduction	Computation Increase
`conv1`	69,696	2,112	33x	33x	133,184	2,112	63x	33x
`conv2`	69,440	1,984	35x	35x	145,600	1,984	73x	35x
`fc1`	17,424	17,424	None	None	17,424	17,424	None	None
`fc2`	36	36	None	None	36	36	None	None

TABLE III
5-FOLD CROSS-VALIDATION RESULT FOR EPILEPTIC SEIZURE DETECTION AND PREDICTION USING OUR NETWORK ARCHITECTURE.

Dataset	Bonn	CHB-MIT					SWEC-ETHZ				
Partition	Set A vs. E	Patient 1	Patient 2	Patient 3	Patient 5	Patient 8	Patient 1	Patient 2	Patient 3	Patient 5	Patient 6
Accuracy	99.84 ± 0.37	99.50 ± 0.89	99.95 ± 0.11	99.95 ± 0.13	99.73 ± 0.57	98.96 ± 2.33	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.86 ± 0.22	100.00 ± 0.00
Sensitivity	99.87 ± 0.28	98.64 ± 2.79	100.00 ± 0.00	100.00 ± 0.00	99.62 ± 0.70	99.76 ± 0.54	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Specificity	99.80 ± 0.45	99.73 ± 0.37	100.00 ± 0.00	99.93 ± 0.15	99.77 ± 0.52	97.38 ± 5.85	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.77 ± 0.39	100.00 ± 0.00
FP per Hour	N/A	0.13 ± 0.17	0.00 ± 0.00	0.03 ± 0.07	0.10 ± 0.22	0.53 ± 1.19	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.13	0.00 ± 0.00
AUROC	99.84 ± 0.37	99.31 ± 1.06	100.00 ± 0.00	99.82 ± 0.39	99.63 ± 0.79	99.04 ± 2.15	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.84 ± 0.25	100.00 ± 0.00

network to be 64, we maintain the flexibility of mapping with scheme (a) to make use of all crossbar rows simultaneously.

As Fig. 6 demonstrates, for parallel convolution layers to be accelerated simultaneously, it was necessary to map the weights of the `conv1` and `conv2` onto two separate crossbar tiles. The weight of the `fc1` layer is a matrix of 1088×8, and using a differential weight scheme, would require 1088×16 memristors. The weight matrix can be further divided into 17 sections of 64×16 weights. To maximize the usage of each 64×64 crossbar array, 4 sections of 64×16 weights can be stacked horizontally onto each crossbar, requiring a total of 5 crossbar tiles.

Since there are unused memristors on the convolution tiles and `fc2` layer operations are not performed immediately after convolution operations, we decided to map the weights of fc2 onto the convolution layer tile, instead of using another tile. Note that since the simulation serves as a validation for proof-of-concept, we decided to use the same dimensions for all 7 crossbar tiles. We do recognize that tile 1, 2 and 7 have many unused memristor devices, as a result, performing small VMMs on a large switch matrix. This leads to large power overhead due to high amortized ADC/DAC power over a small matrix and charge/discharge of long row and column wires without using full length for computation. To address such problem in a real medical device, instead of using square tiles, tile 1, 2 and 7 can be easily mapped onto rectangular tiles of the exact required dimensions.

D. Impact of Device and Crossbar Non-Idealities

Memristors and memristive crossbar arrays are prone to numerous device and circuit non-idealities which have been demonstrated to severely impact the performance of memristive DL accelerators [70]. Consequently, they should be comprehensively simulated prior to circuit-level realization. In this paper, preliminary simulations were performed using the MemTorch [71] simulation framework, and comprehensive simulations of the system using passive crossbar arrays were performed using the crossbar array model provided by [72]. Non-idealities considered include input and output resolutions, weight write resolution, weight write deviation, stuck $R_{\text{ON}}/R_{\text{OFF}}$ devices, line and source resistance, and conductance range variation.

Other memristive phenomena, such as the dynamic behavior of switched memristive neural networks after programming [73], and read disturbance [74], are not accounted for, as practical metal-oxide memristors are endurance-limited, during programming a write-verify scheme is used, and during inference, all Bit Line (BL) voltages are constrained to have a maximum absolute amplitude of 0.3V [74].

E. Stuck Weight Offsetting Methodology

Stuck $R_{\text{ON}}/R_{\text{OFF}}$ weights are known to cause significant network performance degradation in memristive crossbar arrays. Existing works have demonstrated performance recovery through a variety of techniques. In 2014, Kannan et al. took inspiration from SRAM/DRAM technologies and repaired crossbar defects using redundant rows and columns [75]. In 2017, Liu et al. proposed to identify significant weights before applying a retraining and remapping algorithm [76]. In 2018, Xia et al. proposed a mapping algorithm with inner fault tolerance to leverage the differential mapping scheme of crossbar arrays to tolerate faults [77]. In 2019, Zhang et al. proposed the use of matrix transformations to reduce the magnitude of error introduced by stuck-at-fault devices [78]. Also in 2019, Yeo et al. modified conventional transimpedance amplifiers to detect when abnormal current is detected at a particular column due to stuck-at-fault devices and repair by retraining the network with the known defects [79]. Among those works, significant hardware or software overhead is introduced through rewriting and tuning of weights, retraining of networks or using additional circuitry.

To minimize the overhead, we propose stuck weight offsetting, which improves upon the inner fault tolerance method. Inner fault tolerance first identifies all available (non stuck-at-fault) devices and initializes them to default values. Then, the scheme goes through all available devices and adjusts each

TABLE IV
COMPARISON OF OUR BASELINE SOFTWARE MODEL AGAINST SOTA FOR SEIZURE DETECTION USING THE UNIVERSITY OF BONN DATASET

Paper	Pre-processing	Method	Parallelization	Parameters	Accuracy (%)
Ullah *et al.* (2018)	✓	1D-CNN	✗	21,436	99.90
We *et al.* (2018)	✓	1D-CNN	✗	16,778,144	92.00
Abdelhameed *et al.* (2018)	✓	2D-CNN	✗	106,388	98.00
Liu *et al.* (2019)	✓	2D-CNN	✗	N/R*	99.60
Turk *et al.* (2019)	✓	2D-CNN	✗	1,603,080	99.45
Abdelhameed *et al.* (2021)	✓	2D-CNN	✗	10,304,467	**100.00**
Ours	✗	1D-CNN	✓	**10,778**	99.84

*Not reported.

value such that the represented values cannot be made any closer to the target matrix parameter. Intuitively, this serves to minimize the incorrect contribution of the R_{ON}/R_{OFF} weight. We propose to bypass the initialization of available devices to default values and to focus on the complementary weight of stuck-at-fault devices only. Before writing any weights to the crossbar, all stuck-at-fault devices are identified. For each stuck-at-fault device, if the complementary weight is not stuck-at-fault, we calculate its complementary weight to minimize the difference between represented value and target value. All calculated values, along with normal weights, are then written onto the crossbar. This modification reduces overhead by two means. First, all crossbar weights are only required to be written once, as opposed to twice in the inner fault tolerance method (from default to adjusted). Second, our method focuses on complementary weights for stuck-at-fault devices only, as opposed to all available devices for all target parameters. This method incurs minimum additional computational cost, and does not require retraining.

F. Quantization Aware Training for Lower Resolution Systems

A high resolution system is often not feasible to deploy on edge devices, given power consumption constraints and sampling frequency requirements, which are fundamental tradeoffs for resolution in DACs and ADCs. However, lower resolution systems with improved power and frequency performance can exhibit performance degradation. This effect was observed for some patients, and more details can be found in Section VI-C. For significant performance degradation (a degradation of 5% or more compared to full resolution system), we propose to perform Quantization Aware Training (QAT) prior to mapping the weights onto memristive crossbar arrays [86]. During QAT, we quantized the convolutional and fully connected layers of the network to the resolution equivalent to or even lower than that of the resolution of the crossbar weights and ADC/DAC resolution. Quantized layers are implemented using the Brevitas library [86], which provides PyTorch-compatible convolution and fully connected layers of specified weight resolutions. In addition, inputs to the network were quantized, while intermediate outputs remained not quantized. Network architecture and other training parameters remained unchanged.

VI. RESULTS AND DISCUSSION

Prior to the investigation of device and crossbar non-idealities, we report baseline software results for epileptic seizure detection and prediction using our network architecture, in Table III. 5-fold cross-validation was performed using a different seed to eliminate bias on the first fold. To demonstrate the generalizability of the designed network to different domains and patients, the same architecture was applied for seizure detection and prediction. Unlike the Bonn dataset, both the CHB-MIT and SWEC-ETHZ datasets are multi-channel EEG datasets with larger memory and computation requirements within the time domain. In order to reduce the time and memory complexity, pre-processing steps as described in Section IV-A2 were applied to transform the dataset into frequency domain. The shown results suggest that the proposed network is sufficient and can generalize well for both detection and prediction.

A. Comparisons Against SOTA Software Implementations

In Tables IV and V, we compare our baseline software implementations that use full precision (32-bit) floating-point parameters against other software implementations in literature for epileptic seizure detection and prediction, respectively. As shown in the Tables, for epileptic seizure detection we achieve SOTA performance in 3/4 criteria, while for prediction we obtain SOTA performance in 3/6 criteria. Specifically, for detection, our network architecture is able to achieve an accuracy of 99.84% across all samples without any pre-processing steps, while requiring only 10,778 parameters. This is ∼2x fewer parameters than the smallest model in [28], which achieved a slightly higher accuracy of 99.90%, while employing various pre-processing steps. Except for the model used in [87], which achieves a 100% accuracy, but requires over 10M parameters, all the other models shown in Table IV, achieve lower accuracy values despite significantly higher number of network parameters.

For epileptic seizure prediction, pre-processing is performed. Across both datasets, our network architecture achieves the highest sensitivity while requiring the fewest number of parameters. We report close specificity and accuracy values to [15], which has also used a 1D-CNN architecture with parallelization, but needs ∼10x more parameters. Finally, we report the highest FPR across both datasets, however, unlike previous works, we performed no post-processing steps, which may cause this. Also, only two out of the nine previous works have reported their FPR, which makes the comparison incomplete. When mapping trained parameters to ideal crossbars with fully analog devices without any device or circuit non-idealities, the same results were achieved.

B. Generalization Between Datasets

To determine whether or not our trained networks have the ability to generalize, we evaluated the performance of networks trained using the CHB-MIT dataset on the SWEC-ETHZ dataset, and vice-versa in Fig. 7. In addition, we report the cross validation accuracy for networks which have been retrained using transfer learning. To perform transfer learning, parameters were frozen for all layers except the last two fully connected layers, and the weights and biases of the last two fully connected layers were re-trained using

TABLE V
COMPARISON AGAINST SOTA FOR SEIZURE PREDICTION USING THE SWEC-ETHZ AND CHB-MIT DATASETS

Paper	Method	Parallelized	Parameters	Sensitivity (%)	Specificity (%)	Accuracy (%)	FPR†
CHB-MIT							
[66]	2D-CNN	✗	N/R◇	81.20	N/R◇	N/R◇	**0.16**
[80] *	2D-CNN	✗	N/R◇	N/R◇	N/R◇	92.00	N/R◇
[81]	2D-CNN	✗	49,560	82.71	88.21	98.19	N/R◇
[82] *	2D-CNN	✗	N/R◇	88.80	88.60	88.70	N/R◇
[83] *	3D-CNN	✗	28,459,615	96.66	99.14	98.33	N/R◇
[84] *	2D-CNN	✗	9,695,012	84.00	99.00	99.00	0.2
[15]	1D-CNN	✓	105,538	95.55	**99.68**	**99.64**	N/R◇
Ours	1D-CNN	✓	**10,778**	**99.24**	98.68	99.01	0.47
SWEC-ETHZ							
[85] *	Ensemble HD	✗	N/R◇	96.38	97.31	96.85	N/R◇
[15]	1D-CNN	✓	105,538	94.57	**99.86**	**99.81**	N/R◇
Ours	1D-CNN	✓	**10,778**	**98.22**	97.02	97.54	**0.99**

*Indicates the results are reported across the entire dataset and patient-wise performance was not reported. †False positive rate (per hour). ◇Not reported.

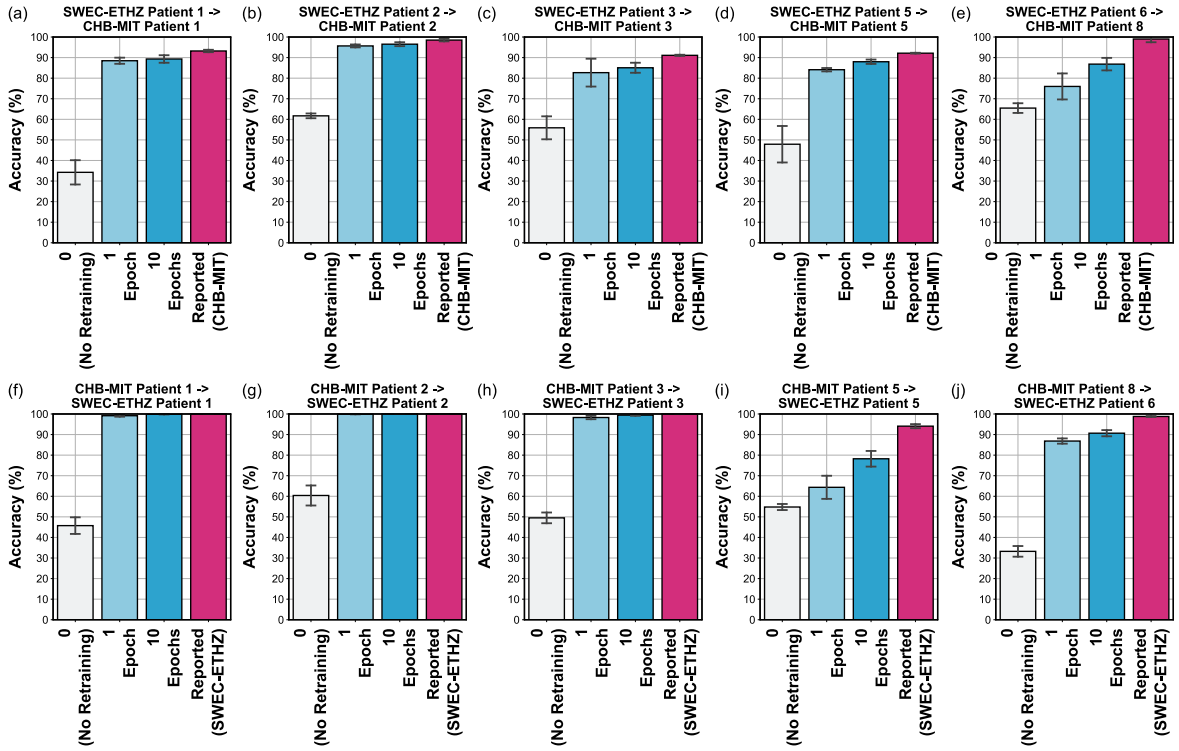


Fig. 7. The ability of our trained networks to generalize between different datasets when performing epileptic seizure prediction. The cross validation accuracy is reported for networks which have not been retrained, and for networks that have been retrained after 1 and 10 training epochs, respectively, when transfer learning was performed. In addition, the standard evaluation accuracy is reported for each dataset and patient, to facilitate comparisons.

the training set of the evaluation dataset. Direct evaluations to/from either of these datasets and the University of Bonn dataset were not made, as the University of Bonn dataset is used for epileptic seizure detection and not prediction, and it is structured differently.

C. Quantization-Aware Training

To demonstrate the effectiveness of QAT, we evaluated the performance of our network architecture when trained with and without QAT. Comparisons are made in Fig. 9. During QAT training, inputs and network weights were reduced to 6-bit resolution, while network architecture and other training parameters were held constant, as described in Fig. 2(f). The accuracy, sensitivity, specificity, AUROC, and FPR metrics were all reported and compared. When using 6-bit ADCs and DACs, it can be observed that for all patients and metrics, except for specificity of patient 5 from the CHB-MIT dataset, QAT network yields significant performance improvements.

D. Effects of Non-Idealities on System Performance

Fig. 8 provides a summary of the impact of non-idealities on our system for epileptic detection and prediction. For the

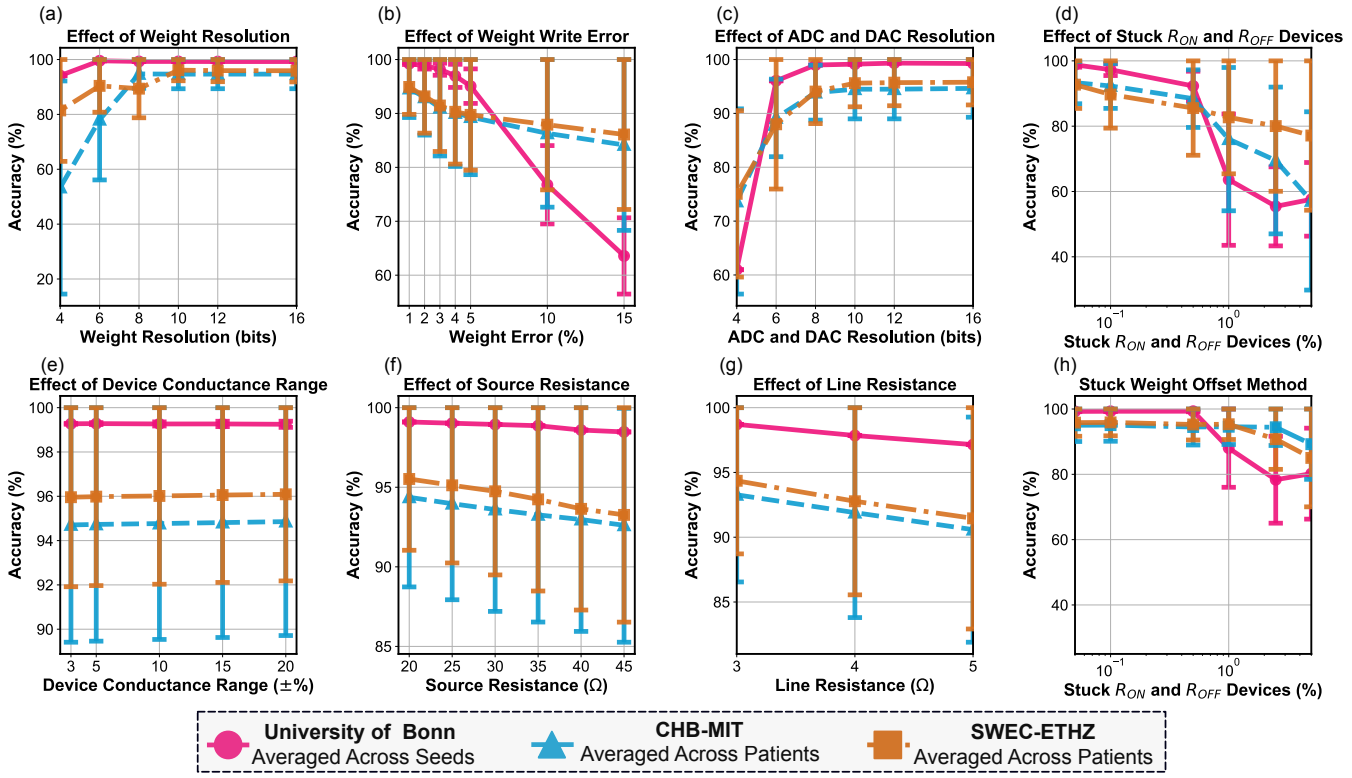


Fig. 8. The impact of all (a-g) non-idealities on the University of Bonn, CHB-MIT, and SWEC-ETHZ datasets. (h) summarizes performance recovery by applying our proposed stuck weight offsetting to address the performance degradation of stuck-at fault devices. For the University of Bonn dataset, each data-point shows the mean and standard deviation across five arbitrary seed values: 5, 6, 7, 8, and 9.

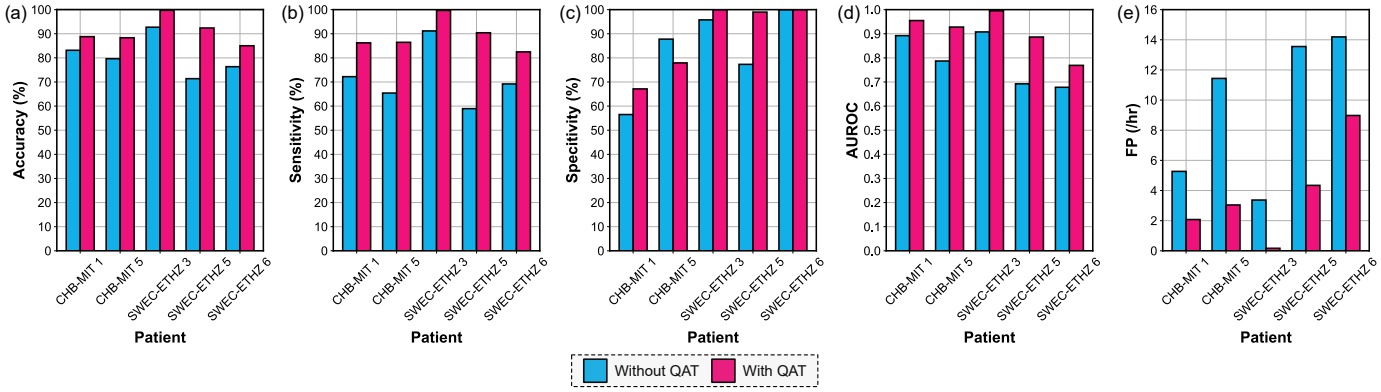


Fig. 9. The impact of QAT on our network architecture tasked for epileptic seizure prediction (a-e) evaluated using the CHB-MIT and SWEC-ETHZ datasets when network parameters are quantized to 6-bit fixed-point resolution. Only patients that exhibited a degradation of 5% or more when quantized to 6-bit fixed-point resolution (from full-precision floating-point) were investigated.

University of Bonn dataset, as samples between patients are not explicitly distinguished, the mean and standard deviation of test set accuracy is reported across samples using five arbitrarily chosen seed values. For the CHB-MIT and SWEC-ETHZ datasets, the mean and standard deviation of test set accuracy is reported across samples for the first five viable patients of each dataset, respectively. Across datasets, some patients were observed to be more robust to non-idealities than others. This was observed in our investigations for patients 1, 2, 3 from the SWEC-ETHZ dataset, and patient 2 from the CHB-MIT dataset, for which non-idealities have minimal impact. For the rest of the patients, however, no clear pattern was established with regards to robustness against non-idealities. We attribute the varying degree of effectiveness between patients to underlying patient specific signatures.

E. Stuck Weight Offsetting

As observed in Fig. 8(d), stuck $R_{\text{ON}}/R_{\text{OFF}}$ devices lead to severe performance degradation. At 1% stuck-at fault and above, system performance can drop below 50% accuracy,

TABLE VI
POWER, AREA, AND LATENCY METRICS FOR THE SIMULATED MEMRISTIVE DL ACCELERATOR USING A 22 NM CMOS PROCESS. USING OUR TDM ARCHITECTURE, VMMS ARE PERFORMED IN $\mathcal{O}(n)$, WHERE n IS THE NUMBER OF COLUMNS OF THE OUTPUT VECTOR. USING OUR PARALLELIZED ARCHITECTURE, VMMS ARE PERFORMED IN $\mathcal{O}(1)$.

Component	Params.	Time-Division Multiplexing (TDM)						Parallelized					
		Specification	Area (mm^2)	Power (mW)	Latency (us)*	Total Latency (us)	Energy (uJ)	Specification	Area (mm^2)	Power (mW)	Latency (us)*	Total Latency (us)	Energy (uJ)
DAC	Resolution Number	6 bits 7x64	2.58E+01	2.69E+03	8.00E-04	2.15E+00	5.78E+00	6 bits 7x64	2.58E+01	2.69E+03	8.00E-04	3.36E-02	9.03E-02
ADC	Resolution Number Frequency	6 bits 7 10MHz	4.62E+00	7.00E+01	1.00E-01	2.69E+02	1.88E+01	6 bits 7x64 10MHz	2.96E+02	4.48E+03	1.00E-01	6.00E-01	2.69E+00
ReLU	Number	2	9.60E-03	3.28E-02	9.80E-02	9.80E-02	3.22E-06	2	9.60E-03	3.28E-02	9.80E-02	9.80E-02	3.22E-06
Average Pool	Number	1	3.83E-04	1.59E+00	8.49E-05	8.49E-05	1.35E-07	1	3.83E-04	1.59E+00	8.49E-05	8.49E-05	1.35E-07
Adder	Number	10	5.34E-03	1.74E-02	3.06E-04	6.13E-04	1.06E-08	10	5.34E-03	1.74E-02	3.06E-04	6.13E-04	1.06E-08
Subtractor	Number	7	2.46E-04	2.87E-01	3.34E-04	1.28E-01	3.69E-05	7x32	7.88E-03	9.20E+00	3.34E-04	2.01E-03	1.85E-05
S+H†	Number	7x64	8.98E-06	3.81E-03	8.33E-04	5.00E-03	1.90E-08	7x64	8.98E-06	3.81E-03	8.33E-04	5.00E-03	1.90E-08
eDRAM Buffer	Size Bus Width	2KB 128	4.72E-03	1.81E+01	1.15E-04	2.30E-04	4.17E-06	2KB 128	4.72E-03	1.81E+01	1.15E-04	2.30E-04	4.17E-06
eDRAM-Tile Bus	Number	192	4.50E-03	3.5E+00	9.02E-05	9.02E-05	3.16E-07	192	4.50E-03	3.5E+00	9.02E-05	9.02E-05	3.16E-07
IR†	Size	1KB	8.10E-01	6.74E-01	8.21E-05	1.64E-04	1.11E-07	1KB	8.10E-01	6.74E-01	8.21E-05	1.64E-04	1.11E-07
OR†	Size	512B	8.70E-04	4.18E-01	8.21E-05	1.64E-04	6.87E-08	512B	8.70E-04	4.18E-01	8.21E-05	1.64E-04	6.87E-08
Scenario: R_{ON}													
Crossbar	Number Size Bits per cell	7 64x64 32	2.87E-04	8.67E+00	2.03E-03	5.82E+01	5.06E-01	7 64x64 32	2.87E-04	8.69E+00	2.03E-03	1.30E-01	1.13E-03
Total			**3.13E+01**	**2.79E+03**		**3.29E+02**	**9.19E+02**		**3.22E+02**	**7.21E+03**		**8.70E-01**	**6.27E+00**
Scenario: $(R_{\text{ON}}^{\sim} + R_{\text{OFF}}^{\sim})/2$													
Crossbar	Number Size Bits per cell	7 64x64 32	2.87E-04	4.35E+00	6.07E-03	1.74E+02	7.58E-01	7 64x64 32	2.87E-04	4.35E+00	6.07E-03	3.88E-01	1.69E-03
Total			**3.13E+01**	**2.79E+03**		**4.45E+02**	**1.24E+03**		**3.22E+02**	**7.21E+03**		**1.13E+00**	**8.12E+00**

*The latency is listed as individual element. †S+H = Sample and Hold, IR = Input Register, OR = Output Register.

rendering the system ineffective. In response to such degradation, we apply our proposed simplified stuck weight offsetting method. Comparing Fig. 8(h) against (d), it is evident that the stuck weight offsetting method improves the average accuracy across all stuck device percentages and datasets. At 1% stuck-at fault, the average accuracy improved by as much as 20% for the Bonn dataset and more than 10% for SWEC-ETHZ and CHB-MIT. The largest improvement was found for the CHB-MIT dataset at 5% stuck-at fault, improving accuracy by 32.11%. At higher stuck device percentages, reduced accuracy recovery is observed. This can be explained by the fact that at higher stuck device percentages, more network information cannot be recovered. Minimizing the contribution of stuck weight cannot fully retrieve the missing information, thereby leading to reduced accuracy recovery. In addition, the proposed method greatly reduces the standard deviation across patients and seeds, thanks to reduced contribution of stuck $R_{\text{ON}}/R_{\text{OFF}}$ devices to final output.

The limitation of this method lies within its inability to deal with both elements of the complementary weight being stuck R_{ON} and R_{OFF} simultaneously. If a positive (negative) weight is stuck R_{ON} and negative (positive) weight is stuck R_{OFF}, stuck weight offsetting cannot provide any further adjustment to minimize the error. Meanwhile, if both weights are stuck R_{ON} or R_{OFF}, the lost weights cannot be recovered, contributing nothing to the final output.

F. Power, Area, and Latency Requirements

The following assumptions, all supported by SOTA DL accelerators, are made when estimating the power, area and latency requirements of our proposed memristive DL accelerator depicted in Fig. 3, targeting a 22nm CMOS process with device integration at the Back-End-Of-The-Line (BEOL). A memristive device has a fixed area of 100×100 nm^2 [103], [104] and the device read latency is 6 ns [105]. An ADC operating frequency is 10 MHz [105], with a power consumption of 10 mW [105] and a device area of 1.1×0.6 mm^2 [104], [106]. A DAC operating frequency is 1.25 GHz, with per unit power consumption of 6 mW and a device area of 0.0576 mm^2 [107]. Other peripheral circuitry with different purposes, including the activation function [108], average pooling layer made up from 4-to-1 multiplexers [109], [110], Sample and Hold (S+H) [111], subtractor [112], and adder [113] circuits, were listed with more detail in Table VI.

All the peripheral components are scaled to 22nm technology by factors introduced in [114] and all buffers with their associated connections have energy, area and latency estimated by CACTI 7.0 [115]. For all calculations, the source resistance and line resistance of 20 Ω and 2 Ω are used respectively. To account for RC delays within crossbars when signals are propagated, the methodology presented in [116] was used, with C_{SA}, $T_{settling}$, and C_{write} parameters from [117]. The largest total device latency was used for all devices.

In Table VI, four scenarios are considered: two where the resistance of all active (utilized) devices was fixed to

TABLE VII
PERFORMANCE SUMMARY AND COMPARISON OF OUR SIMULATED SYSTEM AND EXISTING SEIZURE DETECTION/PREDICTION SYSTEM IMPLEMENTATIONS IN THE LITERATURE.

Paper	Technology	Algorithm(s)	No. Channels	Analog Front-End*	Feature Extract.†	Area (mm^2)	Latency (s)	Power (mW)	Energy (uJ)	Pred.◇	Eval. Task(s)
ML-Based											
[88]	CMOS (180nm)	BPF, LSVM	8	✓	✓	25.00	2.00	N/R°	N/R°	✗	CHB-MIT
[89]	CMOS (180nm)	BPF, NL−SVM	8	✓	✓	25.00	2.00	N/R°	N/R°	✗	CHB-MIT
[90]	CMOS (130nm)	NL−SVM	18	✗	✓	N/R°	4.80	N/R°	N/R°	✗	CHB-MIT
[91]	CMOS (180nm)	FFT, ApEn, LLS	8	✓	✓	13.47	0.8	2.80	2.24E +03	✗	In Vivo
[92]	CMOS (180nm)	BPF, D^2A−LSVM	16	✓	✓	25.0	1.0	N/R°	N/R°	✗	CHB-MIT
[93]	CMOS (180nm)	BPF, NL−SVM	8	✓	✓	25.0	2.0	0.23	460.00	✗	CHB-MIT
[46]	CMOS (130nm)	FIR, PLV	64	✓	✓	3.86	N/R°	1.07	N/R°	✓	In Vivo
[94]	CMOS (130nm)	FIR, PLV/ SE/CFC	32	✓	✓	7.59	0.25	0.71	177.50	✓	In Vivo
[95]	CMOS (180nm)	DWT, KDE, SVM	8	✓	✓	5.83	N/R°	0.67	N/R°	✓	CHB-MIT
[96]	CMOS (40nm)	FFT, NL−SVM	14	✗	✓	4.50	0.71	1.90	1.35E +03	✗	CHB-MIT
[97]	CMOS (65nm)	CHT, XGBoost−DT	16	✓	✓	0.38	N/R°	0.40	N/R°	✗	CHB-MIT, iEEG.org
[98]	CMOS (180nm)	FFT	1	✓	✓	N/R°	N/R°	✗.89	N/R°	✗	CHB-MIT
[99]	CMOS (90nm)	ICA	8	✗	✓	0.4	0.1	8.16E -02	8.16	✗	In Vivo
[72]	CMOS (180nm)	LLS	1	✓	✓	10.41	0.72	2.86E -02	20.59	✗	In Vivo
DL-Based											
[100]	CMOS (65nm)	RNN	8	✗	✗	10.15	N/R°	1✗.80	N/R°	✗	N/R°
[101]	FPGA (M2GL 025-VF256)	MLP	1	✗	✓	N/R°	N/R°	159.70	N/R°	✗	Bonn
[102]	CMOS (180nm)	SNN	1	✗	✓	0.15	64.98E -03	5.40E -03	0.35	✓	In Vivo
Ours (TDM)	**CMOS (22nm)/ RRAM (BEOL)**	**Manual feature extraction, CNN**	**22**	✗	✗	**31.25**	**4.45E -04**	**2.79E +03**	**1.24E +03**	✓	**Bonn, CHB-MIT, ETHZ-SWEC**
Ours (Par.)						**322.31**	**1.13E -06**	**7.20E +03**	**8.12**		

*Reported power, area, and latency requirements include the analog front end/signal acquisition component. †Reported power, area, and latency requirements include feature extraction component(s). ◇Denotes whether systems are able to perform epileptic detection and/or prediction. °Not reported.

$\bar{R_{\text{ON}}} \approx 10\ k\Omega$, while considering either TDM or parallel use of ADC, and two where the average resistance of all active devices was assumed to be $(\bar{R_{\text{ON}}}+\bar{R_{\text{OFF}}})/2 \approx 55\ k\Omega$, again for either TDM or parallelized ADC. These resistance values are representative of two weight distributions: uniform, where all weights are zero, and normal, where all weights are centered around zero. The first distribution was used to report the maximum possible power consumption of our system, and the second distribution was used to report the power consumption of a typical CNN trained using L2-regularization. Considering the marginal impact on total power consumption, (0.16% and 0.06% for TDM and parallelized configurations, respectively), the power of each individual trained CNN was not determined or reported.

For all scenarios, constant operation at 0.3V per cell [74] was assumed. Neither RRAM crossbar tiles nor peripheral circuitry was assumed to be stacked vertically. Consequently, the circuit area consumption was computed as the summation of all individual elements. Both ADCs and DACs were assumed to operate at 6-bit resolution, as stated in Section VI-C, for the best performance with QAT.

As can be observed in Table VI, TDM implementations consume significantly less power than parallelized implementations due to the smaller number of required ADCs. For the worst case TDM scenario, i.e, when all active devices are programmed to $\bar{R_{ON}}$ with a constant 0.3V read voltage, our proposed memristive DL accelerator has a latency of 445.22 μs, and consumes approximately 2.79W and 31.255 mm^2 of power and area. This is fairly low power consumption for a DL accelerator to reside on a separate chip from the neural implant, whereby the implant uses thermal energy to wirelessly communicate with the accelerator [118], for reduced latency.

It is noted that we have chosen to optimize the latency of our system at the cost of higher power consumption for multiple reasons. Firstly, analog crossbars which are used to perform IMC operations, in particular VMMs, require peripheral circuitry which is power- and area-hungry. Consequently, independent of the latency of the system, when inference is being performed, a large proportion of the total system's area and power is consumed by peripheral circuitry, registers, and buffers. While TDM ADCs can be used to reduce the total power consumption by increasing latency, other peripheral circuits, registers, and buffers, are still required for operation. Counterintuitively, in certain instances, the energy of the system can be reduced by minimizing system latency during active operation. In other instances, the performance of the

system can greatly be improved at the cost of increased power consumption.

Secondly, RRAM devices suffer from conductance drift induced by read disturbances, which may aggregate, as the analog current is summed up along each Word Line (WL) during inference [74]. To mitigate this behavior, we have constrained the absolute amplitude of BL voltages to 0.3V and minimized the duration in which a voltage is applied to each device, i.e., latency is minimized to avoid read disturbances, and to prolong the lifespan of RRAM devices, at the cost of increased power consumption. Lastly, as RRAM devices are non-volatile, gating circuitry can be used to reduce the energy consumption of both TDM and parallelized architectures, as both of our architectures have a critical delay path which is much shorter than typical signal acquisition sampling rate periods. This also allows for input buffering to be performed, so that constant operation is not required.

G. Comparison to Existing Hardware Implementations

In Table VII, we compare the performance of hardware implementations of notable epileptic seizure detection and/or prediction hardware systems in the literature. As many different evaluation tasks were used, we did not report performance metrics. Hardware implementations are broadly categorized as either ML- or DL-based. As can be observed, both of our implementations (reported for the $(\bar{R_{\text{ON}}} + \bar{R_{\text{OFF}}})/2$ scenario in Table VI) have significantly reduced inference latency, at the cost of higher power consumption, compared to traditional CMOS and FPGA-based implementations. It is worth noting that, most of the previous designs have not reported a complete power consumption analysis, are not capable of seizure prediction, and use fewer channels, which can lead to lower power consumption and silicon area.

While our proposed system is not currently competitive in resource-constrained environments, it is intended to be used as a reference design for future works implementing epileptic seizure detection and prediction systems using CMOS and memristors. Using analog Static Random-Access Memory (SRAM), vertical stacking of crossbars and CMOS components, and partial sensing approaches, the power and area requirements of our simulated system could be greatly reduced. We aim to investigate these in our future research.

VII. Conclusion

We proposed a parallel CNN architecture that can be used to perform both epileptic seizure detection and prediction rapidly. Compared to other works in literate, our architecture requires significantly fewer parameters, and demonstrates competitive performance on the University of Bonn, CHB-MIT, and SWEC-ETHZ datasets. Using emerging memristive devices and software-hardware optimization methodologies, we demonstrated, through comprehensive simulations, that our memristive DL accelerator is capable of performing real-time operation, and consuming reasonable power in real-world conditions. We also proposed and investigated a new simplified stuck weight offsetting method to improve the robustness of our system to non-idealities. This paper sets a clear path towards the eventual circuit-level realization of a memristive epileptic seizure detection and prediction system.

Acknowledgment

C. Lammie acknowledges the JCU DRTPS and IBM PhD Fellowship Program. M. Rahimi Azghadi acknowledges a JCU Rising Start ECR Fellowship. We thank the handling editor and reviewers' for their constructive feedback. In particular, we acknowledge the second reviewer, who provided advice on simplifying our proposed stuck weight mitigation strategy.

References

[1] E. Beghi, G. Giussani, E. Nichols, F. Abd-Allah, J. Abdela, A. Abdelalim, H. N. Abraha, M. G. Adib, S. Agrawal, F. Alahdab *et al.*, "Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 4, pp. 357–375, 2019.

[2] C. E. Stafstrom and L. Carmant, "Seizures and epilepsy: An overview for neuroscientists," *Cold Spring Harbor Perspectives in Medicine*, vol. 5, no. 6, 2015.

[3] D. C. Patel, B. P. Tewari, L. Chaunsali, and H. Sontheimer, "Neuron-glia interactions in the pathophysiology of epilepsy," *Nature Reviews Neuroscience*, vol. 20, no. 5, pp. 282–297, May 2019.

[4] G. P. Brennan and D. C. Henshall, "micrornas in the pathophysiology of epilepsy," *Neuroscience Letters*, vol. 667, pp. 47–52, 2018, epilepsy: Advances in Genetics and Pathophysiology.

[5] S. Gasparini, E. Ferlazzo, C. Sueri, V. Cianci, M. Ascoli, S. M. Cavalli, E. Beghi, V. Belcastro, A. Bianchi, P. Benna, R. Cantello, D. Consoli, F. A. De Falco, G. Di Gennaro, A. Gambardella, G. L. Gigli, A. Iudice, A. Labate, R. Michelucci, M. Paciaroni, P. Palumbo, A. Primavera, F. Sartucci, P. Striano, F. Villani, E. Russo, G. De Sarro, U. Aguglia, and O. behalf of the Epilepsy Study Group of the Italian Neurological Society, "Hypertension, seizures, and epilepsy: a review on pathophysiology and management," *Neurological Sciences*, vol. 40, no. 9, pp. 1775–1783, Sep. 2019.

[6] M. K. Siddiqui, R. Morales-Menendez, X. Huang, and N. Hussain, "A review of epileptic seizure detection using machine learning classifiers," *Brain informatics*, vol. 7, no. 1, pp. 5–5, May 2020, publisher: Springer Berlin Heidelberg.

[7] F. E. Ibrahim, H. M. Emara, W. El-Shafai, M. Elwekeil, M. Rihan, I. M. Eldokany, T. E. Taha, A. S. El-Fishawy, E.-S. M. El-Rabaie, E. Abdellatef, and F. E. Abd El-Samie, "Deep Learning-based Seizure Detection and Prediction from EEG Signals," *International journal for numerical methods in biomedical engineering*, p. e3573, Jan 2022.

[8] R. Das, "Special Issue on In-Memory Computing," *IEEE Micro*, vol. 42, no. 1, pp. 87–88, 2022.

[9] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, no. 1, pp. 52–59, Jan. 2018.

[10] C. Lammie, O. Krestinskaya, A. James, and M. R. Azghadi, "Variation-aware Binarized Memristive Networks," in *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2019, pp. 490–493.

[11] C. Lammie, W. Xiang, and M. R. Azghadi, "Towards Memristive Deep Learning Systems for Real-Time Mobile Epileptic Seizure Prediction," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.

[13] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," *CoRR*, vol. abs/1605.07146, 2016. [Online]. Available: http://arxiv.org/abs/1605.07146

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. Madison, WI, USA: Omnipress, 2011, p. 689–696.

[15] X. Wang, X. Wang, W. Liu, Z. Chang, T. Kärkkäinen, and F. Cong, "One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg," *Neurocomputing*, vol. 459, pp. 212–222, 2021.

[16] W. Webber, R. P. Lesser, R. T. Richardson, and K. Wilson, "An approach to seizure detection using an artificial neural network (ann)," *Electroencephalography and clinical Neurophysiology*, vol. 98, no. 4, pp. 250–272, 1996.

[17] N. Pradhan, P. Sadasivan, and G. Arunodaya, "Detection of seizure activity in eeg by an artificial neural network: A preliminary study," *Computers and Biomedical Research*, vol. 29, no. 4, pp. 303–313, 1996.

[18] A. M. Chan, F. T. Sun, E. H. Boto, and B. M. Wingeier, "Automated seizure onset detection for accurate onset time determination in intracranial eeg," *Clinical Neurophysiology*, vol. 119, no. 12, pp. 2687–2696, 2008.

[19] T. Netoff, Y. Park, and K. Parhi, "Seizure prediction using cost-sensitive support vector machine," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 3322–3325.

[20] K. Chua, V. Chandran, U. R. Acharya, and C. Lim, "Automatic identification of epileptic electroencephalography signals using higher-order spectra," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 223, no. 4, pp. 485–495, 2009.

[21] T. L. Sorensen, U. L. Olsen, I. Conradsen, J. Henriksen, T. W. Kjaer, C. E. Thomsen, and H. B. Sorensen, "Automatic epileptic seizure onset detection using matching pursuit: a case study," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 3277–3280.

[22] L. Chisci, A. Mavino, G. Perferi, M. Sciandrone, C. Anile, G. Colicchio, and F. Fuggetta, "Real-time epileptic seizure prediction using ar models and support vector machines," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1124–1132, 2010.

[23] E. B. Petersen, J. Duun-Henriksen, A. Mazzaretto, T. W. Kjær, C. E. Thomsen, and H. B. Sorensen, "Generic single-channel detection of absence seizures," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4820–4823.

[24] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "Eeg-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[25] U. R. Acharya, S. V. Sree, and J. S. Suri, "Automatic detection of epileptic eeg signals using higher order cumulant features," *International journal of neural systems*, vol. 21, no. 05, pp. 403–414, 2011.

[26] A. Kharbouch, A. Shoeb, J. Guttag, and S. S. Cash, "An algorithm for seizure onset detection using intracranial eeg," *Epilepsy & Behavior*, vol. 22, pp. S29–S35, 2011.

[27] Y. Liu, W. Zhou, Q. Yuan, and S. Chen, "Automatic seizure detection using wavelet transform and svm in long-term intracranial eeg," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 20, no. 6, pp. 749–755, 2012.

[28] I. Ullah, M. Hussain, E. ul Haq Qazi, and H. Aboalsamh, "An automated system for epilepsy detection using eeg brain signals based on deep learning approach," *Expert Systems with Applications*, vol. 107, pp. 61 – 71, 2018.

[29] W. Zhao, W. Zhao, W. Wang, X. Jiang, X. Zhang, Y. Peng, B. Zhang, and G. Zhang, "A Novel Deep Neural Network for Robust Detection of Seizures Using EEG Signals," *Computational and Mathematical Methods in Medicine*, vol. 2020, p. 9689821, Apr. 2020, publisher: Hindawi.

[30] R. Abiyev, M. Arslan, J. Bush Idoko, B. Sekeroglu, and A. Ilhan, "Identification of epileptic eeg signals using convolutional neural networks," *Applied Sciences*, vol. 10, no. 12, 2020.

[31] P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri, "A comparison of deep neural networks for seizure detection in eeg signals," *bioRxiv*, 2019.

[32] S. Liss, "Method and apparatus for monitoring and counteracting excess brain electrical energy to prevent epileptic seizures and the like," *US3850161A*, 1973.

[33] S. Viglione, V. Ordon, W. Martin, and C. Kesler, "Epileptic seizure warning system," *US3863625A*, 1973.

[34] A. Aarabi and B. He, "A rule-based seizure prediction method for focal neocortical epilepsy," *Clinical Neurophysiology*, vol. 123, no. 6, pp. 1111–1122, 2012.

[35] S. Li, W. Zhou, Q. Yuan, and Y. Liu, "Seizure prediction using spike rate of intracranial eeg," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 21, no. 6, pp. 880–886, 2013.

[36] A. S. Zandi, R. Tafreshi, M. Javidan, and G. A. Dumont, "Predicting epileptic seizures in scalp eeg based on a variational bayesian gaussian mixture model of zero-crossing intervals," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1401–1413, 2013.

[37] M. Bedeeuzzaman, T. Fathima, Y. U. Khan, and O. Farooq, "Seizure prediction using statistical dispersion measures of intracranial eeg," *Biomedical Signal Processing and Control*, vol. 10, pp. 338–341, 2014.

[38] P. E. McSharry, T. He, L. A. Smith, and L. Tarassenko, "Linear and non-linear methods for automatic seizure detection in scalp electroencephalogram recordings," *Medical and Biological Engineering and Computing*, vol. 40, no. 4, pp. 447–461, 2002.

[39] B. Schelter, H. Feldwisch-Drentrup, M. Ihle, A. Schulze-Bonhage, and J. Timmer, "Seizure prediction in epilepsy: From circadian concepts via probabilistic forecasting to statistical evaluation," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1624–1627.

[40] S. Wang, W. A. Chaovalitwongse, and S. Wong, "A novel reinforcement learning framework for online adaptive seizure prediction," in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2010, pp. 499–504.

[41] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 804–813, 2019.

[42] R. Jana and I. Mukherjee, "Deep learning based efficient epileptic seizure prediction with eeg channel optimization," *Biomedical Signal Processing and Control*, vol. 68, p. 102767, 2021.

[43] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Patient-independent epileptic seizure prediction using deep learning models," *arXiv preprint arXiv:2011.09581*, 2020.

[44] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. Abd El-Samie, "Eeg seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 183, Dec. 2014.

[45] M. R. Azghadi, C. Lammie, J. K. Eshraghian, M. Payvand, E. Donati, B. Linares-Barranco, and G. Indiveri, "Hardware implementation of deep network accelerators towards healthcare and biomedical applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 6, pp. 1138–1159, 2020.

[46] H. Kassiri, S. Tonekaboni, M. T. Salam, N. Soltani, K. Abdelhalim, J. L. P. Velazquez, and R. Genov, "Closed-loop neurostimulators: A survey and a seizure-predicting design example for intractable epilepsy treatment," *IEEE transactions on biomedical circuits and systems*, vol. 11, no. 5, pp. 1026–1040, 2017.

[47] D. Kudithipudi, Q. Saleh, C. Merkel, J. Thesing, and B. Wysocki, "Design and analysis of a neuromemristive reservoir computing architecture for biosignal processing," *Frontiers in Neuroscience*, vol. 9, p. 502, 2016.

[48] C. Merkel, Q. Saleh, C. Donahue, and D. Kudithipudi, "Memristive reservoir computing architecture for epileptic seizure detection," *Procedia Computer Science*, vol. 41, pp. 249 – 254, 2014, 5th Annual International Conference on Biologically Inspired Cognitive Architectures, 2014 BICA.

[49] R.-E. Karamani, I.-A. Fyrigos, V. Ntinas, I. Vourkas, G. C. Sirakoulis, and A. Rubio, "Memristive cellular automata for modeling of epileptic brain activity," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[50] Z. Liu, J. Tang, B. Gao, P. Yao, X. Li, D. Liu, Y. Zhou, H. Qian, B. Hong, and H. Wu, "Neural signal analysis with memristor arrays towards high-efficiency brain–machine interfaces," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[51] S. M. Usman, M. Usman, and S. Fong, "Epileptic seizures prediction using machine learning methods," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[52] K. Fujiwara, M. Miyajima, T. Yamakawa, E. Abe, Y. Suzuki, Y. Sawada, M. Kano, T. Maehara, K. Ohta, T. Sasai-Sakuma *et al.*, "Epileptic seizure prediction based on multivariate statistical process control of heart rate variability features," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1321–1332, 2015.

[53] M. Zanghieri, A. Burrello, S. Benatti, K. Schindler, and L. Benini, "Low-latency detection of epileptic seizures from ieeg with temporal convolutional networks on a low-power parallel mcu," in *2021 IEEE Sensors Applications Symposium (SAS)*. IEEE, 2021, pp. 1–6.

[54] C. N. Heck, D. King-Stephens, A. D. Massey, D. R. Nair, B. C. Jobst, G. L. Barkley, V. Salanova, A. J. Cole, M. C. Smith, R. P. Gwinn *et al.*, "Two-year seizure reduction in adults with medically intractable partial onset epilepsy treated with responsive neurostimulation: final

results of the rns system pivotal trial," *Epilepsia*, vol. 55, no. 3, pp. 432–441, 2014.

[55] M. Nasseri, T. Pal Attia, B. Joseph, N. M. Gregg, E. S. Nurse, P. F. Viana, G. Worrell, M. Dümpelmann, M. P. Richardson, D. R. Freestone *et al.*, "Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[56] Y. Yang, M. Zhou, Y. Niu, C. Li, R. Cao, B. Wang, P. Yan, Y. Ma, and J. Xiang, "Epileptic seizure prediction based on permutation entropy," *Frontiers in computational neuroscience*, vol. 12, p. 55, 2018.

[57] M. Pinto, A. Leal, F. Lopes, A. Dourado, P. Martins, C. A. Teixeira *et al.*, "A personalized and evolutionary algorithm for interpretable eeg epilepsy seizure prediction," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[58] R. E. Stirling, D. B. Grayden, W. D'Souza, M. J. Cook, E. Nurse, D. R. Freestone, D. E. Payne, B. H. Brinkmann, T. Pal Attia, P. F. Viana *et al.*, "Forecasting seizure likelihood with wearable technology," *Frontiers in neurology*, p. 1170, 2021.

[59] P. Peng, Y. Song, and L. Yang, "Seizure prediction in eeg signals using stft and domain adaptation," *Frontiers in Neuroscience*, p. 1880, 2021.

[60] J. Cong and B. Xiao, "Minimizing computation in convolutional neural networks," in *Artificial Neural Networks and Machine Learning – ICANN 2014*, S. Wermter, C. Weber, W. Duch, T. Honkela, P. Koprinkova-Hristova, S. Magg, G. Palm, and A. E. P. Villa, Eds. Cham: Springer International Publishing, 2014, pp. 281–290.

[61] A. Amirsoleimani, F. Alibart, V. Yon, J. Xu, M. R. Pazhouhandeh, S. Ecoffey, Y. Beilliard, R. Genov, and D. Drouin, "In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits: Design choices, challenges, and perspectives," *Advanced Intelligent Systems*, vol. 2, no. 11, p. 2000115, 2020.

[62] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov, "4k-memristor analog-grade passive crossbar circuit," *Nature Communications*, vol. 12, no. 1, p. 5198, Aug. 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-25455-0

[63] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state." *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 64, p. 061907, Dec 2001.

[64] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, Aug. 2021.

[65] A. Burrello, L. Cavigelli, K. Schindler, L. Benini, and A. Rahimi, "Laelaps: An energy-efficient seizure detection algorithm from long-term human ieeg recordings without false alarms," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2019, pp. 752–757.

[66] N. D. Truong, A. D. Nguyen, L. Kuhlmann, M. R. Bonyadi, J. Yang, S. Ippolito, and O. Kavehei, "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Networks*, vol. 105, pp. 104–111, Sep. 2018.

[67] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, Jul. 2020. [Online]. Available: https://doi.org/10.1038/s41565-020-0655-z

[68] M. Rahimi Azghadi, Y.-C. Chen, J. K. Eshraghian, J. Chen, C.-Y. Lin, A. Amirsoleimani, A. Mehonic, A. J. Kenyon, B. Fowler, J. C. Lee, and Y.-F. Chang, "Complementary metal-oxide semiconductor and memristive hardware for neuromorphic computing," *Advanced Intelligent Systems*, vol. 2, no. 5, p. 1900189, 2020.

[69] K. Chellapilla, S. Puri, and P. Simard, "High Performance Convolutional Neural Networks for Document Processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, G. Lorette, Ed., Université de Rennes 1. La Baule (France): Suvisoft, Oct. 2006.

[70] O. Krestinskaya, A. Irmanova, and A. P. James, "Memristive non-idealities: Is there any practical implications for designing neural network chips?" in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.

[71] C. Lammie, W. Xiang, B. Linares-Barranco, and M. Rahimi Azghadi, "MemTorch: An Open-source Simulation Framework for Memristive Deep Learning Systems," *Neurocomputing*, vol. 485, pp. 124–133, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222002053

[72] A. Chen, "A comprehensive crossbar array model with solutions for line resistance and nonlinear device characteristics," *IEEE Transactions on Electron Devices*, vol. 60, no. 4, pp. 1318–1326, 2013.

[73] J. Cheng, L. Liang, J. H. Park, H. Yan, and K. Li, "A Dynamic Event-Triggered Approach to State Estimation for Switched Memristive Neural Networks With Nonhomogeneous Sojourn Probabilities," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 12, pp. 4924–4934, 2021.

[74] W. Shim, Y. Luo, J.-s. Seo, and S. Yu, "Impact of Read Disturb on Multilevel RRAM based Inference Engine: Experiments and Model Prediction," in *2020 IEEE International Reliability Physics Symposium (IRPS)*, 2020, pp. 1–5.

[75] S. Kannan, N. Karimi, R. Karri, and O. Sinanoglu, "Detection, diagnosis, and repair of faults in memristor-based memories," in *2014 IEEE 32nd VLSI Test Symposium (VTS)*. IEEE, 2014, pp. 1–6.

[76] C. Liu, M. Hu, J. P. Strachan, and H. Li, "Rescuing memristor-based neuromorphic design with high defects," in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2017, pp. 1–6.

[77] L. Xia, W. Huangfu, T. Tang, X. Yin, K. Chakrabarty, Y. Xie, Y. Wang, and H. Yang, "Stuck-at fault tolerance in rram computing systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 102–115, 2017.

[78] B. Zhang, N. Uysal, D. Fan, and R. Ewetz, "Handling stuck-at-faults in memristor crossbar arrays using matrix transformations," in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, 2019, pp. 438–443.

[79] I. Yeo, M. Chu, S.-G. Gi, H. Hwang, and B.-G. Lee, "Stuck-at-fault tolerant schemes for memristor crossbar array-based neural networks," *IEEE Transactions on Electron Devices*, vol. 66, no. 7, pp. 2937–2945, 2019.

[80] T. Wen and Z. Zhang, "Deep convolution neural network and autoencoders-based unsupervised feature learning of eeg signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018.

[81] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–17, 2019.

[82] J. Cao, J. Zhu, W. Hu, and A. Kummert, "Epileptic signal classification with deep eeg features by stacked cnns," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 4, pp. 709–722, 2019.

[83] X. Tian, Z. Deng, W. Ying, K.-S. Choi, D. Wu, B. Qin, J. Wang, H. Shen, and S. Wang, "Deep multi-view feature learning for eeg-based epileptic seizure detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 1962–1972, 2019.

[84] W. Liang, H. Pei, Q. Cai, and Y. Wang, "Scalp eeg epileptogenic zone recognition and localization based on long-term recurrent convolutional network," *Neurocomputing*, vol. 396, pp. 569–576, 2020.

[85] A. Burrello, S. Benatti, K. Schindler, L. Benini, and A. Rahimi, "An ensemble of hyperdimensional classifiers: Hardware-friendly short-latency seizure detection with automatic ieeg electrode selection," *IEEE journal of biomedical and health informatics*, vol. 25, no. 4, pp. 935–946, 2020.

[86] A. Pappalardo, "Xilinx/brevitas," 2021.

[87] A. Abdelhameed and M. Bayoumi, "A deep learning approach for automatic seizure detection in children with epilepsy," *Frontiers in Computational Neuroscience*, vol. 15, p. 29, 2021.

[88] J. Yoo, L. Yan, D. El-Damak, M. A. B. Altaf, A. H. Shoeb, and A. P. Chandrakasan, "An 8-channel scalable eeg acquisition soc with patient-specific seizure classification and recording processor," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 214–228, 2013.

[89] M. A. B. Altaf, J. Tillak, Y. Kifle, and J. Yoo, "A $1.83/micro$j/classification nonlinear support-vector-machine-based patient-specific seizure classification soc," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2013, pp. 100–101.

[90] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, 2013.

[91] W.-M. Chen, H. Chiueh, T.-J. Chen, C.-L. Ho, C. Jeng, M.-D. Ker, C.-Y. Lin, Y.-C. Huang, C.-W. Chou, T.-Y. Fan, M.-S. Cheng, Y.-L. Hsin, S.-F. Liang, Y.-L. Wang, F.-Z. Shaw, Y.-H. Huang, C.-H. Yang, and C.-Y. Wu, "A fully integrated 8-channel closed-loop neural-prosthetic

cmos soc for real-time epileptic seizure control," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 232–247, 2014.

[92] M. A. Bin Altaf, C. Zhang, and J. Yoo, "A 16-channel patient-specific seizure onset and termination detection soc with impedance-adaptive transcranial electrical stimulator," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 11, pp. 2728–2740, 2015.

[93] M. A. Bin Altaf and J. Yoo, "A 1.83μj/classification, 8-channel, patient-specific epileptic seizure classification soc using a non-linear support vector machine," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 1, pp. 49–60, 2016.

[94] G. O'Leary, M. R. Pazhouhandeh, M. Chang, D. Groppe, T. A. Valiante, N. Verma, and R. Genov, "A recursive-memory brain-state classifier with 32-channel track-and-zoom Δ2 $\sum$ADCs and Charge-Balanced Programmable Waveform Neurostimulators," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*.

[95] Y. Wang, Q. Sun, H. Luo, X. Chen, X. Wang, and H. Zhang, "26.3 a closed-loop neuromodulation chipset with 2-level classification achieving 1.5v¡inf¿pp¡/inf¿ cm interference tolerance, 35db stimulation artifact rejection in 0.5ms and 97.8IEEE International Solid- State Circuits Conference - (ISSCC), 2020, pp. 406–408.

[96] S.-A. Huang, K.-C. Chang, H.-H. Liou, and C.-H. Yang, "A 1.9-mw svm processor with on-chip active learning for epileptic seizure control," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 2, pp. 452–464, 2020.

[97] A. Uran, K. Ture, C. Aprile, A. Trouillet, F. Fallegger, A. Emami, S. P. Lacour, C. Dehollain, Y. Leblebici, and V. Cevher, "A 16-channel wireless neural recording system-on-chip with cht feature extraction processor in 65nm cmos," in *2021 IEEE Custom Integrated Circuits Conference (CICC)*, 2021, pp. 1–2.

[98] S.-K. Lin, Istiqomah, L.-C. Wang, C.-Y. Lin, and H. Chiueh, "An ultra-low power smart headband for real-time epileptic seizure detection," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–10, 2018.

[99] C.-H. Yang, Y.-H. Shih, and H. Chiueh, "An 81.6 μW FastICA Processor for Epileptic Seizure Detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 1, pp. 60–71, 2015.

[100] C. Chen, H. Ding, H. Peng, H. Zhu, Y. Wang, and C.-J. R. Shi, "Ocean: An on-chip incremental-learning enhanced artificial neural network processor with multiple gated-recurrent-unit accelerators," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 519–530, 2018.

[101] H. G. Daoud, A. M. Abdelhameed, and M. Bayoumi, "Fpga implementation of high accuracy automatic epileptic seizure detection system," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018, pp. 407–410.

[102] M. Ronchini, M. Zamani, H. A. Huynh, Y. Rezaeiyan, G. Panuccio, H. Farkhani, and F. Moradi, "A CMOS-based neuromorphic device for seizure detection from LFP signals," *Journal of Physics D: Applied Physics*, vol. 55, no. 1, p. 014001, oct 2021.

[103] S. Lv, J. Liu, and Z. Geng, "Application of memristors in hardware security: A current state-of-the-art technology," *Advanced Intelligent Systems*, vol. 3, no. 1, p. 2000127, 2021.

[104] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31–56, 2021.

[105] M. A. Zidan, Y. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, and W. D. Lu, "A general memristor-based partial differential equation solver," *Nature Electronics*, vol. 1, no. 7, pp. 411–420, Jul. 2018.

[106] M. Yoshioka, M. Kudo, K. Gotoh, and Y. Watanabe, "A 10 b 125 ms/s 40 mw pipelined adc in 0.18 /spl mu/m cmos," in *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, 2005, pp. 282–598 Vol. 1.

[107] J. Jung, K.-H. Baek, S.-I. Lim, S. Kim, and S.-M. Kang, "Design of a 6 bit 1.25 gs/s dac for wpan," in *2008 IEEE International Symposium on Circuits and Systems*, 2008, pp. 2262–2265.

[108] M. Giordano, G. Cristiano, K. Ishibashi, S. Ambrogio, H. Tsai, G. W. Burr, and P. Narayanan, "Analog-to-digital conversion with reconfigurable function mapping for neural networks activation function acceleration," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 367–376, 2019.

[109] Z. Li, A. Ren, J. Li, Q. Qiu, Y. Wang, and B. Yuan, "Dscnn: Hardware-oriented optimization for stochastic computing based deep convolutional neural networks," in *2016 IEEE 34th International Conference on Computer Design (ICCD)*, 2016, pp. 678–681.

[110] P.-E. Gaillardon, M. H. Ben-Jamaa, F. Clermidy, and I. O'Connor, "Evaluation of a crossbar multiplexer in a lithography-based nanowire technology," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, 2011, pp. 2930–2933.

[111] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 2016, pp. 14–26.

[112] K. Govindarajan and V. S. K. Bhaaskaran, "Borrow Select Subtractor for Low Power and Area Efficiency," in *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2020, pp. 518–523.

[113] S. Ganesan *et al.*, "Area, delay and power comparison of adder topologies," Ph.D. dissertation, 2015.

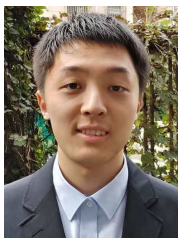
[114] S. Sarangi and B. Baas, "Deepscaletool: A tool for the accurate estimation of technology scaling in the deep-submicron era," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[115] R. Balasubramonian, A. B. Kahng, N. Muralimanohar, A. Shafiee, and V. Srinivas, "Cacti 7: New tools for interconnect exploration in innovative off-chip memories," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 2, jun 2017.

[116] A. Dozortsev, I. Goldshtein, and S. Kvatinsky, "Analysis of the row grounding technique in a memristor-based crossbar array," *International Journal of Circuit Theory and Applications*, vol. 46, no. 1, pp. 122–137, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cta.2399

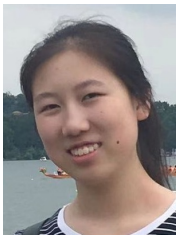
[117] C. Xu, X. Dong, N. P. Jouppi, and Y. Xie, "Design implications of memristor-based rram cross-point structures," in *2011 Design, Automation Test in Europe*, 2011, pp. 1–6.

[118] M. ElAnsary, J. Xu, J. Sales Filho, G. Dutta, L. Long, C. Tejeiro, A. Shoukry, C. Tang, E. Kilinc, J. Joshi *et al.*, "Bidirectional peripheral nerve interface with 64 second-order opamp-less adcs and fully integrated wireless power/data transmission," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 11, pp. 3247–3262, 2021.

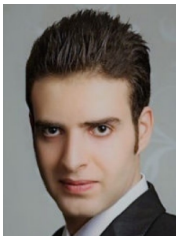


Chenqi Li is currently pursuing a B.A.Sc in Engineering Science, Robotics at University of Toronto, Canada. His current research interests include machine learning, computer vision, and brain-inspired computing.

Corey Lammie (S'17) is currently pursuing a PhD in Computer Engineering at James Cook University (JCU), where he completed his undergraduate degrees in Electrical Engineering (Honours) and Information Technology in 2018. His main research interests include brain-inspired computing, and the simulation and hardware implementation of Spiking Neural Networks (SNNs) and Artificial Neural Networks (ANNs) using RRAM devices and FPGAs. He has received several awards and fellowships including the intensely competitive 2020-2021 IBM international PhD Fellowship, a Domestic Prestige Research Training Program Scholarship (the highest paid PhD scholarship in Australia), the 2020 Circuits and Systems (CAS) Society Pre-Doctoral Grant, and the 2017 Engineers Australia CN Barton Medal awarded to the best undergraduate engineering thesis at JCU. Corey has served as a reviewer for several IEEE journals and conferences including IEEE Transactions on Circuits and Systems I and II, and the IEEE International Symposium on Circuits and Systems (ISCAS).



Xuening Dong is currently pursuing a B.A.Sc in Computer Engineering at University of Toronto, Canada. Her current research interests include machine learning, stochastic processes, and the design and simulation of memristor-based applications.



Amirali Amirsoleimani (S'09–M'2017) is an assistant professor in the Department of Electrical Engineering and Computer Science at the Lassonde School of Engineering. He received his PhD in electrical and computer engineering (ECE) from University of Windsor in December 2017 and completed his postdoctoral research fellowship at the Edward S. Rogers Sr. Electrical and Computer Engineering Department at the University of Toronto in July 2021. His current research interests include application-specific processing units, in-memory computing, neuromorphic hardware design and RRAM-based accelerators for artificial intelligence. He received IEEE Larry K. Wilson award for IEEE region 7 in 2016. He was also the recipient of a best poster honourable mention award at International Joint Conference on Neural Network (IJCNN) 2017 in Alaska, USA. He is a guest editor in Frontiers in Electronics and Frontiers in Nanotechnology journals and is also serving as a reviewer for several electrical and computer engineering journals including IEEE Transactions on Circuits and Systems I (TCASI), TCAS II, TNANO, TVLSI, TED, Frontiers in Neuro-Science, Microelectronics journal, Neural Computing and Applications.

Mostafa Rahimi Azghadi (S'07–M'14–SM'19) completed his PhD in Electrical Electronic Engineering at The University of Adelaide, Australia, earning the Doctoral Research Medal, as well as the Adelaide University Alumni Medal. He is currently a senior lecturer in the College of Science and Engineering, James Cook University, Townsville, Australia, where he researches low-power and high-performance neuromorphic accelerators for neural inspired and deep learning networks for a variety of applications from agriculture to medicine. He has co-raised over $6M in research funding from national and international resources.

Dr. Rahimi was a recipient of several national and international accolades including a 2015 South Australia Science Excellence award, a 2016 Endeavour Research Fellowship, a 2017 Queensland Young Tall Poppy Science Award, a 2018 JCU Rising Star ECR Leader Fellowship, a 2019 Fresh Science Queensland Finalist, and a 2020 JCU Award for Excellence in Innovation and Change. Dr Rahimi is a senior member of the IEEE and a TC member of Neural Systems and Applications of the circuit and system society. He serves as an associate editor of Frontiers in Neuromorphic Engineering and IEEE Access.

Roman Genov (S'96–M'02–SM'11) received the B.S. degree in Electrical Engineering from Rochester Institute of Technology, NY in 1996 and the M.S.E. and Ph.D. degrees in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, MD in 1998 and 2003 respectively.

He is currently a Professor in the Department of Electrical and Computer Engineering at the University of Toronto, Canada, where he is a member of Electronics Group and Biomedical Engineering Group and the Director of Intelligent Sensory Microsystems Laboratory. Dr. Genov's research interests are primarily in analog integrated circuits and systems for energy-constrained biological, medical, and consumer sensory applications.

Dr. Genov is a co-recipient of Jack Kilby Award for Outstanding Student Paper at IEEE International Solid-State Circuits Conference, Best Paper Award of IEEE Transactions on Biomedical Circuits and Systems, Best Paper Award of IEEE Biomedical Circuits and Systems Conference, Best Student Paper Award of IEEE International Symposium on Circuits and Systems, Best Paper Award of IEEE Circuits and Systems Society Sensory Systems Technical Committee, Brian L. Barge Award for Excellence in Microsystems Integration, MEMSCAP Microsystems Design Award, DALSA Corporation Award for Excellence in Microsystems Innovation, and Canadian Institutes of Health Research Next Generation Award. He was a Technical Program Co-chair at IEEE Biomedical Circuits and Systems Conference, a member of IEEE European Solid-State Circuits Conference Technical Program Committee, and a member of IEEE International Solid-State Circuits Conference International Program Committee. He was also an Associate Editor of IEEE TCAS II and IEEE Signal Processing Letters, as well as a Guest Editor for IEEE JSSC. Currently he is an Associate Editor of IEEE Transactions on Biomedical Circuits and Systems.