Customized Development and Hardware Optimization of a Fully-Spiking SNN for EEG-Based Seizure Detection

Abdul Muneeb, Hossein Kassiri
Department of Electrical Engineering and Computer Science, York University, Toronto, Canada

2024 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-5495-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BioCAS61083.2024.10798216

***Abstract*—We present a customized approach in development of a fully-spiking neural networks (SNNs) tailored for energy-efficient data-driven signal processing in implantable brain neural interfaces. The importance of a customized design lies in its ability to optimize hardware and energy efficiency while maintaining high classification performance. Our approach allows for the customization of key parameters, including quantization resolution of weights and biases, encoding scheme, encoder placement, temporal resolution, neuron type, and internal parameters such as threshold value, resetting process, and refractory period. We demonstrated the efficacy of this customization in improving hardware and energy efficiency through model development, software-based training and testing, and subsequent synthesis using Verilog RTL on FPGAs and ASIC implementation. Performance evaluation using the CHB-MIT dataset showed an average sensitivity of 92.2% and specificity of 97.3% for seizure detection. The synthesis reports provided insights into the memory, computation, and energy requirements for hardware implementation, highlighting the efficiency and effectiveness of our approach. Our results show that SNN models leads to only 1% drop of sensitivity compared with a 32-bit real-value resolution SCNN model, while offering more than 4 times improvement in memory efficiency.**

***Index Terms*– spiking neural network (SNN), electroencephalogram (EEG), seizure detection, neuromorphic, energy-efficient, computational cost, epilepsy.**

I. Introduction

Energy-efficient data-driven signal processing for implantable brain neural interfaces is of paramount importance, particularly in applications such as seizure detection, which require real-time analysis of multi-channels recordings while constrained by stringent implant size and battery life [1]–[6]. This has made the need for effective low-power implementation of such algorithms increasingly critical [7]–[10]. Spiking neural networks (SNNs) have emerged as a promising approach for developing such energy-efficient systems [12]–[16]. SNNs offer distinct advantages over deep neural networks (DNNs) due to their asynchronous operation and spike-based computation. The asynchronous operation allows them to efficiently leverage neural signals' information sparsity, processing data only on-demand and in response to input events. The spike-based computation replaces conventional costly multi-bit binary multiply-and-accumulate (MAC) operations with inexpensive single-bit accumulations, thus reducing power/resource consumption.

Currently, there are two options available for developing SNNs, each with its own set of advantages and challenges. Organic SNN training, which directly trains the network using spike-based learning rules, can capture the temporal dynamics of neural activity more naturally, leading to potentially more biologically plausible models. However, it can be challenging due to the complex and non-differentiable nature of spike-timing-dependent plasticity [17]–[19]. In contrast, shadow training, which involves training a DNN and then converting the model to a spiking one, leverages the well-established training algorithms of DNNs. This approach generally results in higher accuracy and more stable performance, as it benefits from the robust optimization techniques available for DNNs before conversion to an SNN [20]–[22].

The DNN-to-SNN conversion in shadow training involves several critical decisions, including the number of time steps, spiking threshold values, encoding scheme, and quantization resolution of weights and biases. Another significant aspect is ensuring all layers of the network are of spiking nature (i.e., spike encoding is done prior to the network and not within the network), particularly the first layer, which handles a substantial portion of the computations. As will be discussed in detail, the currently-available toolboxes and libraries do not fully capitalize on the advantages of spike-based computation as they convert DNNs to partially-spiking neural networks and offer limited customization options for spike encoding, weight/bias quantization, and temporal resolution.

In this work, we employ a fully customizable approach for the development of SNNs, which streamlines model conversion from Python to hardware description language (HDL)-based FPGA implementation, followed by VLSI synthesis for ASIC implementation. By offering comprehensive customization of encoding schemes, weight and bias quantization resolution, and other critical parameters, our framework optimizes both energy and area efficiency of the SNN's hardware implementation. We will discuss various design choices throughout the development, conversion, and hardware implementation processes, and present simulation results showcasing the performance of the proposed approach in terms of seizure detection accuracy, latency, and energy efficiency.

II. Motivation for a Customizable SNN Shadow Training

Considering the event-driven asynchronous nature of SNNs, it is reasonable to assume that the number of spikes decrease rapidly as we move forward in the network. Assuming an exponential decrease, the first layer processes a significant majority of the operations. To estimate the first layer's computational quota in an SNN with n layers, each having N neurons per layer and an exponential spike decay factor r ($0 < r < 1$), let T_i be the number of spikes per neuron in layer i, with $T_{i+1} = r \cdot T_i$. The total computations in the network, C_{total}, and the computations in the first layer, C_1, are given by:

$$C_{\text{total}} = \sum_{i=1}^{n} N \times T_i = N \times T_1 \left(1 + r + r^2 + \cdots + r^{n-1}\right) \quad (1)$$

$$C_{\text{total}} = N \times T_1 \frac{1 - r^n}{1 - r} \quad (2)$$

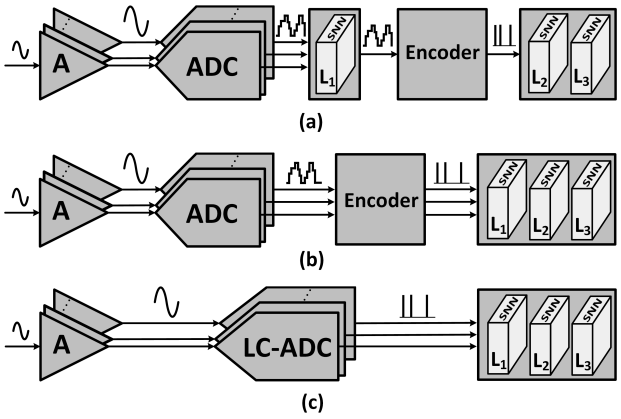

Fig. 1. (a) A partially-spiking neural network (PSNN) compared to (b) and fully-spiking neural network (FSNN) realized through pre-SNN encoding. (c) An FSNN interfaced with asynchronous recording circuit.

The first layer quota is (FLQ):

$$\text{FLQ} = \frac{C_1}{C_{\text{total}}} = \frac{N \times T_1}{N \times T_1 \frac{1-r^n}{1-r}} = \frac{1-r}{1-r^n} \tag{3}$$

This estimation underscores the importance of optimizing the first layer for efficient spike processing, as it will significantly influence the overall performance and energy efficiency of the SNN. Moreover as shown in Fig. 1, a fully-spiking SNN (i.e., including the first layer) also allows for integrating them with event-driven, asynchronous spiking front-ends, which can unlock significant power savings for implantable brain-computer interfaces (BCIs). Traditional analog front-ends with amplifiers and ADCs consume substantial power, diminishing the benefits of an energy-efficient neuromorphic processor. In comparison to conventional front ends, Level crossing ADCs (LC-ADCs) respond to variations in the input signal in an asynchronous fashion, resulting in event-driven power consumption and data output [23].

Various toolboxes/libraries are publicly available for SNN development and training. For instance, SNN Toolbox [22] is a versatile tool designed to facilitate the conversion of trained ANNs to SNNs as shown in Fig. 2. It supports various deep learning frameworks, enabling seamless integration and conversion. Another example is SNN Torch [19], which is a library built on top of PyTorch, providing a comprehensive framework for creating, training, and evaluating SNNs. Despite their advantages, these tools have critical constraints (listed in Table I), which prevent from optimizing the SNN in terms of computational and energy efficiency.

TABLE I. CONSTRAINTS OF SNN DEVELOPMENT APPROACHES

Implementation	Hardware	Fully Spiking	Quantization	Encoder before first layer
Spiking-CNN [30]	×	✓	×	✓
Spiking-CNN [21]	✓	×	×	✓
SNN Toolbox [22]	×	×	×	×
SNN Torch [19]	×	×	×	×
Spiking Conformer [15]	×	×	×	×

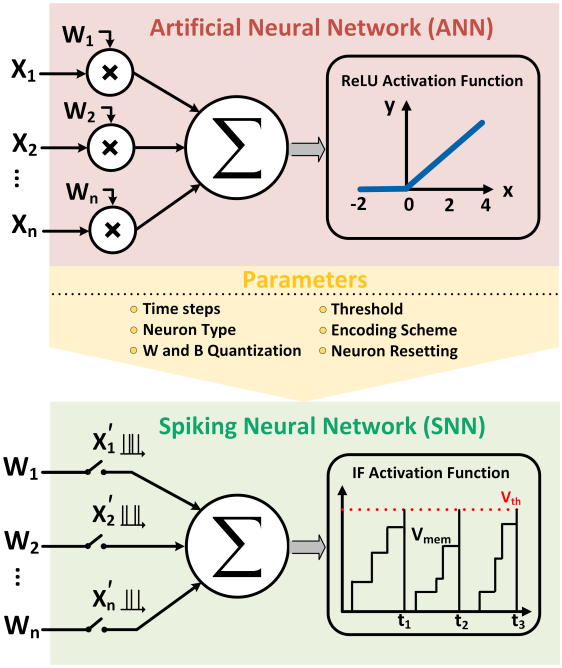

Fig. 2. High-level visualization of the shadow training process for SNN development and the critical parameters that need to be decided to ensure a hardware-optimized implementation.

III. MODEL DEVELOPMENT AND EVALUATION

A. CNN Design and Conversion Methodology

Fig. 3 shows the top-level architecture of the proposed seizure detection system, in which the presented spiking convolutional neural network (SCNN) is connected to the output of an array of LC-ADCs, which feeds asynchronous temporally-coded data to the network. For SCNN architecture design (i.e., numbers of convolutional layers, size of kernels, etc.) we used neural architecture search (NAS), which involves automatically generating and evaluating numerous candidate architectures, assessing their performance, and iteratively refining the search process [24]. The performance of proposed model was evaluated through training, cross-validation, and testing phases using a labeled scalp EEG dataset [25]. The final design has four convolutional layers, including three temporal-convolution layers to extract time-domain features from different channels, followed by one spatial-convolution layer to extract cross-channel spatial features from the input data. To enhance training and mitigate the model overfitting problem, each convolutional layer is followed by batch normalization and max pooling. Following all these layers three fully connected layers are implemented.

Next, we converted the developed CNN to a spiking CNN using a custom-made process. Instead of relying on built-in libraries for SNN inference, we implemented the entire system from scratch, utilizing only Numpy. Our approach was based on a theoretical understanding of SNN operations, including the internal mechanisms of spiking neurons, layer interactions, and overall network function. We first developed a single spiking neuron, defining its internal mechanism for asynchronous membrane potential updates, including integration and resetting. We then defined layers, kernels, and the convolution process in a fully customizable, parametric fashion, ultimately constructing the network with a customizable number of layers. This approach allowed for choosing all aspects of the resulted SCNN, including network architecture (i.e., number of layers,

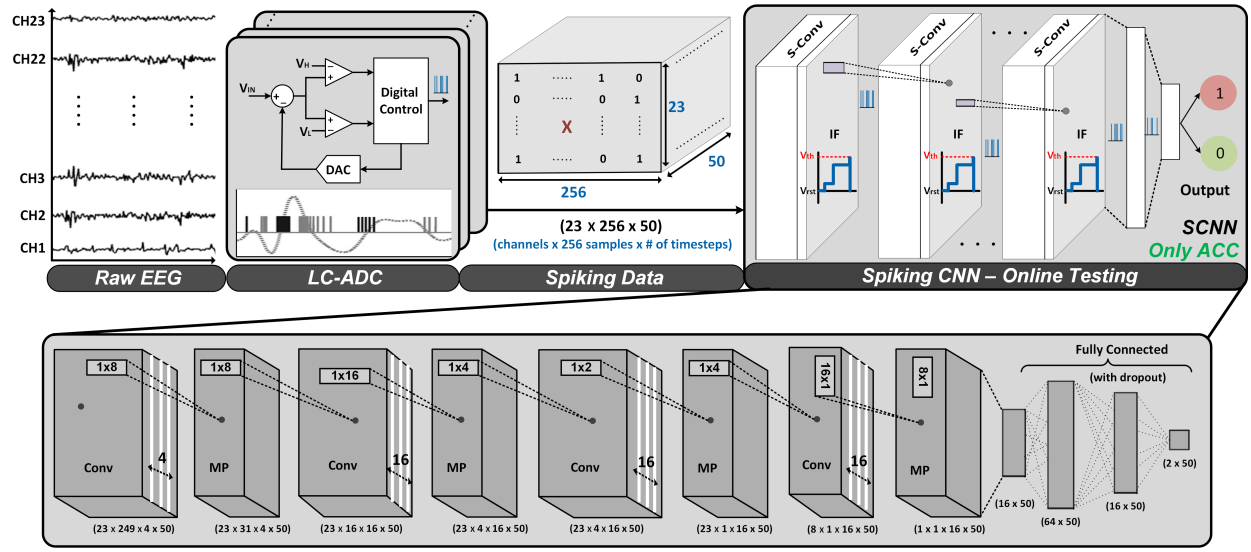

Fig. 3. Top-level block diagram of the system architecture, illustrating the characteristics of the input signals and encoded spiking data. The diagram also details the architecture of the spiking convolutional neural network, which utilizes 1D temporal and spatial kernels. All kernel and layer dimensions are shown for clarity.

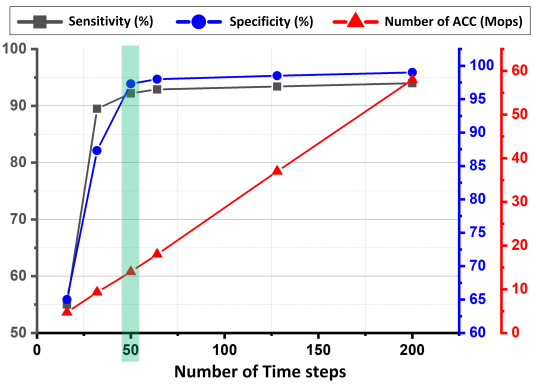

Fig. 4. Seizure detection sensitivity, specificity, and computational cost as a function of the SNN's temporal resolution.

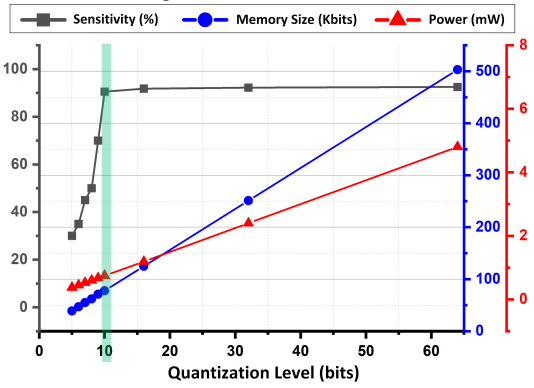

Fig. 5. Seizure detection sensitivity, SNN model size, and power consumption as a function of quantization resolution used for weights and bias values.

neurons per layer, and connectivity patterns (fully connected, convolutional, etc)), neuron model, encoding scheme, and parameters quantization resolution. We also ensured that we have visibility to all nodes of the network and can monitor signals received and generated by each neurons, as well as their real-time membrane potential.

B. Temporal Encoding and Resolution

As illustrated in Fig. 2, in addition to weights and biases that are decided through CNN training, the CNN to SCNN

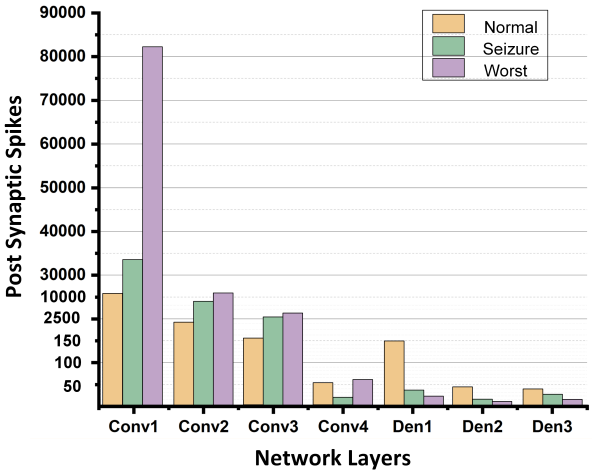

Fig. 6. Number of Post synaptic spikes at the output of each layer of the network in different situations.

conversion requires making decision on encoding scheme and temporal resolution. In this work, we used Poisson rate encoding due to its straightforward and intuitive process, as well as compatibility with LC-ADCs. We chose the number of time-steps, which is equivalent to number of time-domain quantization levels, by striking a trade-off between computational efficiency and seizure detection accuracy. Fig. 4 shows the performance evaluation and the trade-off between number of time steps and computational cost for the developed SNN architecture, suggesting 50 to be an optimal number, beyond which detection performance does not improve significantly.

C. Neuron Model

We use Integrate and fire (IF) neuron model, for which the discrete-time equation for updating the membrane potential is given by

$$V(t+1) = V(t) + I(t) \tag{4}$$

where $V(t)$ is the membrane potential at time t and $I(t)$ is the input current at time t. If the updated membrane potential $V(t+1)$ exceeds the firing threshold θ, a spike is generated and

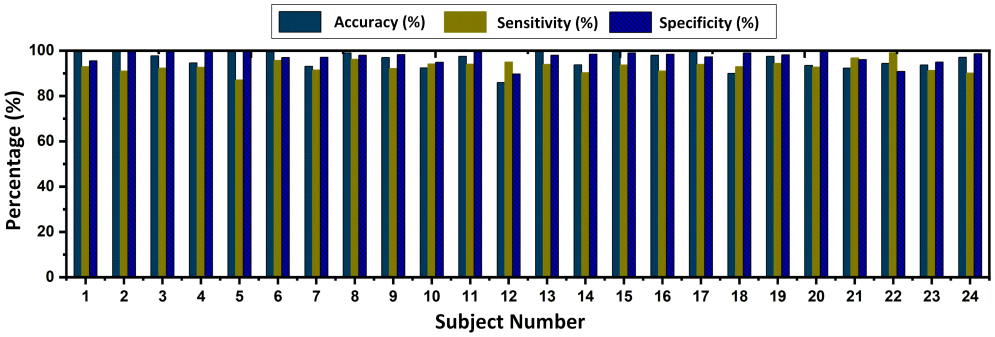

Fig. 7. Subject-wise seizure detection performance of the custom-developed SCNN.

the membrane potential is reset by subtracting the threshold:

$$\text{if } V(t+1) \geq \theta, \text{ then } V(t+1) = V(t+1) - \theta \qquad (5)$$

To avoid immature spike generation instead of resetting to zero, we reset membrane potential through subtraction of the threshold value. To further improve energy and computational efficiency of the IF neuron model, a gating mechanism was embedded in it to prune zero-value post-synaptic spikes.

D. Parameters Quantization

While the values for weights and biases are decided through training, the quantization resolution used for storing and conducting computations with them is a critical aspect of network's hardware implementation. Fig. 5 shows the design tradeoff for deciding the quantization resolution. As shown, although the weights and biases found through training process are initially 32-bit floating-point numbers, reducing the resolution down to 10-bit fixed-point value seem to have an insignificant impact on seizure detection performance. This is in line with what we anticipated considering (a) the dynamic range of weight/bias values, and (b) the input signals' quantization resolution is in the same range as well, making higher resolution for the weights an unnecessary overdesign.

E. Model Testing

The open-source CHB-MIT dataset was used for training and testing. It contains scalp EEG data from 24 patients ranging in age from 1.5 to 22 years, sampled at 256Hz [25]. Fig. 6 shows the average number of spikes for ten seconds of data at the output of each layer during normal periods, seizure episodes, and the worst case scenario, i.e., with the most active input dynamics. The data confirms that as spikes propagate through the network, they generally become sparser as discussed earlier, leading to reduced computational costs in intermediate and deeper layers. It also illustrates the importance of a fully-spiking network and conducting spike encoding before the first layer, considering its share of computations compared to other layers.

For each patient, we used 60% of data for training, 10% for cross validation, and 30% for testing. At this stage, we only focused on sensitivity and specificity of seizure detection, as defined in [26]. For performance evaluation, we used vector-based metrics, as defined in [27]. The subject wise performance metrics for the SCNN are presented in Fig. 7, which show an average sensitivity and specificity of 92.2% and 97.3%, respectively both for the SCNN.

TABLE II. SCNN PARAMETERS AND SYNTHESIS RESULTS

Parameters		Synthesis Report	
Memory	9.75KB	Process	Standard 0.13 μm CMOS
Cell count	78.9K	Voltage	1.8 V
Time Steps	50	Total Power	1.37mW
Neuron Model	Integrate and Fire	Total Area	$3.3mm^2$
Thresholding	Thresh: 1, Reset by subtraction	Quantization Resolution	10-bits

TABLE III. COMPARISON WITH THE STATE-OF-THE-ART PATIENT-SPECIFIC SEIZURE DETECTION ALGORITHMS

Ref		BSPC'23 [31]	ISCAS'24 [15]	TBME'22 [32]	JSSC'22 [27]	This Work	
Classifier		SNN+2DLSVM	Spiking Conformer	CNN	GTCA-SVM	CNN	SCNN
HW Imp		ASIC	ASIC	N/A	ASIC	STM32H747	ASIC
No. of Parameter		14.9K	9.9K	N/A	N/A	7.8K	7.8K
Model Size		59.6KB	39.6KB	45.2KB	70KB	31.2KB	9.75KB
No. of Operations		6K ACC	0.32M ADD, 1.0K MUL	-	N/A	0.65M ACC, 0.66M MUL	$1.5M^{\circledast}$, $0.2M^{\diamond}$ ACC
Feature Extraction		Power Spectrum	Raw EEG	N/A	Raw EEG	Raw EEG	Raw EEG
Sensitivity (%)	Vector	N/R	N/R	N/R	97.8	93.4	92.2
	Event	88.4	94.9	93.4	100	100	100
Specificity/FPR		84.6%	99.3%	0.27/hr	99.5%	99.1%	97.3%
Time steps		200	8	N/R	N/R	N/A	50

N/R: Not Reported, N/A: Not Applicable
MUL: Multiplication, ACC: Accumulation $\circledast$: Avg. for Seizure, $\diamond$: Avg. for Non-seizure

IV. HARDWARE IMPLEMENTATION

We converted the developed SCNN to hardware description language (Verilog RTL) and then synthesized using a standard 130nm CMOS cells library. The synthesis report from Cadence Genus is shown in Table II. Table III compares the presented fully spiking SCNN with the developed SCNN and CNN architectures from the state of the art, highlighting the excellent overall performance of the presented design.

V. CONCLUSION

In this work, we presented a fully customized approach to developing fully-spiking neural networks (SNNs) for energy-efficient data-driven signal processing in implantable brain neural interfaces. Our approach allows for comprehensive customization of key parameters including the quantization resolution of weights and biases, encoding scheme, encoder placement, temporal resolution (i.e., the number of time steps), neuron type, and its internal parameters such as threshold value, resetting process, and refractory period.

Through the development process, we have highlighted various design choices and their impacts on the performance of the SNN. Following software-based training and testing, we synthesized the SNN using Verilog RTL on FPGA and subsequently for ASIC implementation. The synthesis reports provided insights into the memory, computation, and energy requirements for hardware implementation.

Performance evaluation using the CHB-MIT dataset demonstrated an average sensitivity of 92.2% and specificity of 97.3% for seizure detection. These results underscore the efficiency and effectiveness of our customized SNN approach, while significantly improving hardware efficiency.

REFERENCES

[1] A. Muneeb, M. Ali and M. A. B. Altaf, "A 2.7μJ/classification Machine-Learning based Approximate Computing Seizure Detection SoC," *IEEE International Symposium on Circuits and Systems (ISCAS),* May, 2022.

[2] T. Zhan, et al., "A Resource-Optimized VLSI Architecture for Patient-Specific Seizure Detection using Frontal-Lobe EEG,"*IEEE International Symposium on Circuits and Systems (ISCAS),* 2019, pp. 1-5.

[3] H. Kassiri, N. Soltani, M. T. Salam, J. L. Perez Velazquez and R. Genov, "Battery-less modular responsive neurostimulator for prediction and abortion of epileptic seizures,"*IEEE International Symposium on Circuits and Systems (ISCAS),* 2016, pp. 1298-1301.

[4] M. Bialer, S. I. Johannessen, R. H. Levy, E. Perucca, T. Tomson, H. S. White, and M. J. Koepp, "Seizure detection and neuromodulation: A summary of data presented at the XIII conference on new antiepileptic drugs and devices (EILAT XIII),"*Epilepsy Res.,* vol. 130, pp. 27-36, 2017.

[5] A. Dabbaghian, et al., "Modular Flexible 80-dB-DR Artifact-Resilient EEG Headset with Distributed Pulse-Based Feature Extraction and Multiplier-Less Neuromorphic Boosted Seizure Classifier,"*IEEE Custom Integrated Circuits Conference (CICC),* 2024, pp. 1-2.

[6] M. T. Salam et al., "Tradeoffs between wireless communication and computation in closed-loop implantable devices,"*IEEE International Symposium on Circuits and Systems (ISCAS),* 2016, pp. 1838-1841.

[7] H. Kassiri et al., "Battery-less tri-band-radio neuro-monitor and responsive neurostimulator for diagnostics and treatment of neurological disorders,"*IEEE J. Solid-State Circuits,* vol. 51, no. 5, pp. 1274-1289, 2016.

[8] M. R. Karimi et al., "A multi-feature nonlinear-SVM seizure detection algorithm with patient-specific channel selection and feature customization,"*IEEE International Symposium on Circuits and Systems (ISCAS),* 2020, pp. 1-5.

[9] T. Zhan et al., "A resource-optimized VLSI implementation of a patient-specific seizure detection algorithm on a custom-made 2.2 cm ^{2} wireless device for ambulatory epilepsy diagnostics,"*IEEE Trans. Biomed. Circuits Syst.,* vol. 13, no. 6, pp. 1175-1185, 2019.

[10] A. Muneeb et al., "Energy-Efficient Spiking-CNN-Based Cross-Patient Seizure Detection,"*IEEE Int. Symp. Circuits Syst. (ISCAS),* 2023, pp. 1-5.

[11] H. Kassiri et al., "Closed-loop neurostimulators: A survey and a seizure-predicting design example for intractable epilepsy treatment,"*IEEE Trans. Biomed. Circuits Syst.,* vol. 11, no. 5, pp. 1026-1040, 2017.

[12] K. Roy, et al., "Towards spike-based machine intelligence with neuromorphic computing,"*Nature,* 2019.

[13] K. Yamazaki, et al., "Spiking neural networks and their applications: A review,"*Brain Sciences,* 2022.

[14] A. Bhattacharjee, et al., "Are SNNs Truly Energy-efficient? — A Hardware Perspective,"*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2024.

[15] Q. Chen, et al., "Epilepsy Seizure Detection and Prediction using an Approximate Spiking Convolutional Transformer,"*IEEE International Symposium on Circuits and Systems (ISCAS),* 2024.

[16] J. Wang, S. Zhao, J. Yang and M. Sawan, "An Event-driven Neural Signal Processor for Closed-loop Seizure Prediction,"*IEEE Biomedical Circuits and Systems Conference (BioCAS),* 2023.

[17] C. Sun et al., "An Energy Efficient STDP-Based SNN Architecture With On-Chip Learning,"*IEEE Transactions on Circuits and Systems,* vol. 69, no. 12, pp. 5147-5158, Dec. 2022.

[18] Q. Chen, X. Dong, D. Ma and X. Zhu, "A Hardware Accelerator of the Convolutional Spike Neural Network Based on STDP Online Learning,"*Conference of Science and Technology for Integrated Circuits (CSTIC),* 2024.

[19] J. K. Eshraghian et al., "Training Spiking Neural Networks Using Lessons From Deep Learning," *Proc. of the IEEE,* vol. 111, no. 9, pp. 1016-1054, Sept. 2023.

[20] S. Hwang and J. Kung, "One-Spike SNN: Single-Spike Phase Coding With Base Manipulation for ANN-to-SNN Conversion Loss Minimization,"*IEEE Transactions on Emerging Topics in Computing,* 2024.

[21] A. Muneeb, S. Mehrotra and H. Kassiri, "A 9.5ms-Latency 6.2μJ/Inference Spiking CNN for Patient-Specific Seizure Detection," *IEEE Biomedical Circuits and Systems Conference (BioCAS),* Oct. 2023.

[22] B. Rueckauer, et al., "Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification," *Frontiers in Neuroscience,* vol. 11, Dec. 2017.

[23] J. V. Assche and G. Gielen, "Analysis and Design of a 10.4-ENOB 0.92–5.38- μ W Event-Driven Level-Crossing ADC With Adaptive Clocking for Time-Sparse Edge Applications," *IEEE Journal of Solid-State Circuits,* Mar. 2024.

[24] K. T. Venkata, and A. K. Somani, "Neural Architecture Search Survey: A Hardware Perspective," *ACM Comput. Survey,* vol. 55, no. 4, pp. 1000-1036, November 2022.

[25] A. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment,," *Ph.D. dissertation, Massachusetts Inst. Technol.,* Cambridge, MA, USA, Sep. 2009.

[26] M. Salam, et al., "Rapid brief feedback intracerebral stimulation based on real-time desynchronization detection preceding seizures stops the generation of convulsive paroxysms." Epilepsia, vol. 56, no. 8 pp. 1227-1238, 2015.

[27] M. Zhang, L. Zhang, C. -W. Tsai and J. Yoo, "A Patient-Specific Closed-Loop Epilepsy Management SoC With One-Shot Learning and Online Tuning," *IEEE J. Solid-State Circuits,* vol. 57, no. 4, pp. 1049-1060, April 2022.

[28] S. Kim, et al., "C-DNN: An Energy-Efficient Complementary Deep-Neural-Network Processor With Heterogeneous CNN/SNN Core Architecture,," *IEEE Journal of Solid-State Circuits,* vol. 59, no. 1, pp. 157-172, Jan. 2024.

[29] C. W. Tsai, et al., "SciCNN: A 0-Shot-Retraining Patient-Independent Epilepsy-Tracking SoC," *IEEE International Solid- State Circuits Conference (ISSCC),* 2023.

[30] F. Tian, J. Yang, S. Zhao and M. Sawan, "A New Neuromorphic Computing Approach for Epileptic Seizure Prediction," *IEEE International Symposium on Circuits and Systems (ISCAS),* 2021.

[31] H. Shan, et al., "Compact seizure detection based on spiking neural network and support vector machine for efficient neuromorphic implementation," *Biomedical Signal Processing and Control (BSPC),* vol. 86, p. 105268, 2023.

[32] Shiqi Zhao , Jie Yang , and Mohamad Sawan, "Energy-Efficient Neural Network for Epileptic Seizure Prediction," *IEEE Trans. on Biomedical Engineering,* vol. 69, no. 1, 2022.