Low-Power and Low-Cost AI Processor with Distributed-Aggregated Classification Architecture for Wearable Epilepsy Seizure Detection

Qiang Zhang, Mingyue Cui, Yue Liu, Weichong Chen, Zhiyi Yu, *Senior Member, IEEE,*

***Abstract*—Wearable devices with continuous monitoring capabilities are critical for the daily detection of epileptic seizures, as they provide users with accurate and comprehensible analytical results. However, current AI classifiers rely on a two-stage recognition process for continuous monitoring, which only reduces operation time but remains challenged by the high cost of additional hardware. To address this problem, this article proposes a novel fusion architecture for AI processors, which enables event-triggered cross-paradigm integration and computation. Our method introduces a distributed-aggregated classification architecture (D-ACA) that facilitates the reuse of hardware resources across two-stage recognition, thereby obviating the need for standby hardware and enhancing energy efficiency. Integrating a non-encoding biomedical circuit method based on spiking neural networks (SNNs), the architecture eliminates encoded neurons at the hardware level, significantly optimizing energy consumption and hardware resource utilization. Additionally, we develop a configurable and highly flexible control method that supports various neuron modules, enabling continuous detection of epileptic seizures and activating high-precision recognition upon event detection. Finally, we implement the design on the Xilinx ZCU 102 FPGA board, where the AI processor achieves a high classification accuracy of 98.1% while consuming extremely low classification energy (3.73 μJ per classification).**

***Index Terms*—epilepsy seizure, EEG signal processing, low-power and low-cost, processor, spiking neural networks**

I. Introduction

Epilepsy seizure detection is increasingly gaining attention in the current healthcare field, as it is a highly debilitating and painful condition that significantly impacts patients' lives. Timely detection and intervention are the best methods for managing epileptic seizures. However, heavy and non-wearable devices fail to meet the requirements for real-time monitoring and mobility, causing significant inconvenience to patients. Designing a wearable detection device that is both low-power and low-cost, while ensuring accurate recognition, is of critical importance.

Currently, wearable intelligent biomedical signal devices for detecting epileptic seizures garner significant attention due to their potential for enabling home health monitoring and automatic anomaly detection [1]–[4]. Compared to traditional seizure detection devices, these advanced systems integrate artificial intelligence (AI) analytical capabilities to identify anomalies in biomedical signals and alert users for further medical evaluation [5]–[9]. AI processors used for detecting epileptic seizures are crucial components of these devices, enabling intelligent classification of electroencephalogram (EEG) signals into normal and abnormal patterns. However, the design of AI processors presents several challenges, particularly the need for ultra-low power consumption, stringent accuracy requirements, and biomedical adaptability [10].

The development of AI processors is primarily aimed at accelerating neural networks for general AI applications, such as image and speech recognition [11]. However, these processors are not well-suited for biomedical AI processing, which necessitates the integration of biomedical and AI processing hardware. Fortunately, the methods proposed in [2], [12]–[14] address the requirements of biomedical neural signal processing and support AI hardware. However, these methods fail to address the unique characteristics of epilepsy, including the following: 1) The short interval between seizures necessitates lightweight devices with long standby times; 2) The short duration of seizures requires detection devices to respond precisely and perform accurate recognition. Consequently, these methods do not fundamentally provide novel solutions and pose challenges in achieving high performance and low power consumption, both of which are crucial in biomedical applications.

There are also several biomedical AI processors, such as those designed for electrocardiogram (ECG), EEG, and electromyogram (EMG) signal processing [12], [15], [16]. The redundancy inherent in their design for general AI applications means that their precise recognition classifiers are constantly active, leading to significant power consumption and rendering them unsuitable for long-term operation in epilepsy seizure detection devices. Currently, researchers are exploring solutions to this issue by leveraging the characteristic of extremely high sparsity in epileptic seizures. Previous work [2] employs an event-driven method for seizure detection, utilizing a bio-inspired spiking neural network. This processor maintains continuous detection with low power consumption and switches to a high-accuracy convolutional neural network (CNN)-based recognition mode only upon event detection, thereby enhancing versatility and energy efficiency. Additionally, references [12], [13] introduce a two-level hardware

Qiang Zhang, Weichong Chen, and Yue Liu are with the School of Microelectronics Science and Technology, Sun Yat-Sen University, Zhuhai, 519082, China. Email: zhangq553@mail2.sysu.edu.cn, chenwch36@mail2.sysu.edu.cn, liuy2389@mail2.sysu.edu.cn.

Zhiyi Yu is with the School of Microelectronics Science and Technology, Guangdong Provincial Key Laboratory of Optoelectronic Information Processing Chips and Systems, Sun Yat-sen University, Zhuhai, 519082, China. Email: yuzhiyi@mail.sysu.edu.cn (Corresponding author: Zhiyi Yu).

Mingyue Cui is with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 400100, China. Email: cuimy@mail2.sysu.edu.cn.

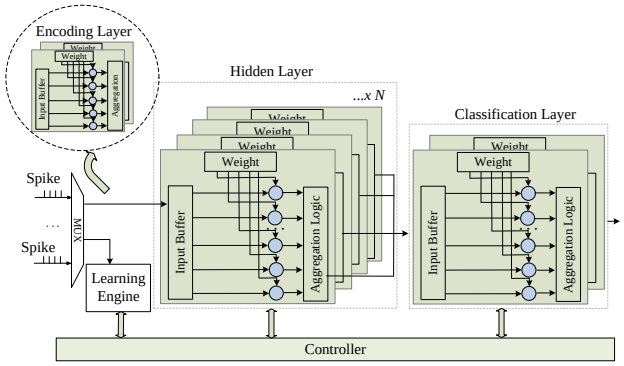


Fig. 1: Potential simplification of the processing flow for biomedical and AI hardware

computing architecture that similarly achieves the desired energy efficiency. However, they unavoidably rely on assistance from standby hardware. These two-stage processing methods essentially entail the integration of hardware of varying scales, resulting in a heterogeneous architecture. As discussed in [2], [12], [13], in most cases, the precise classifier of the second level consumes a significant portion of hardware resources but remains idle most of the time. This underutilization of hardware resources contradicts the design philosophy of ultra-low-power wearable devices, making it a significant challenge to address.

In our work, we propose an AI processor for epilepsy seizure detection that incorporates several advanced design techniques. The distributed-aggregated classification architecture (D-ACA) offers a method to reconfigure the architecture through an event-triggered mechanism. D-ACA allows for the reuse of hardware in a two-level recognition process, thereby eliminating the need for standby hardware. To the best of our knowledge, our work represents the first attempt to eliminate the encoding components of traditional neural network architectures in hardware circuits when processing spike-based bioelectric signals. This potential simplification streamlines the processing flow, reducing hardware consumption and power consumption, as shown in Fig. 1. We compare our method against methods such as HNNA [2], CLIC [17], MDMR [1], BioAIP [18], and IoMT [19] in the experimental section. The processor achieves low power consumption per classification (3.73 μJ/classification) and high classification accuracy (98.1%) with an SE of 95.28%, along with at least a 3.46$\times$ reduction in memory usage and a 1.68$\times$ reduction in hardware resource consumption. We make an effective trade-off between addressing the aforementioned issues and maintaining accuracy, aiming to optimize performance, enhance energy efficiency, and reduce hardware resource consumption. The technical details are as follows:

- The D-ACA provides a method for reconfiguring the architecture through event-triggered mechanisms. D-ACA enables hardware reuse across two-stage recognition, eliminating standby hardware. Additionally, the processor adopts a modular design and circuit-level integration, achieving neuron fusion at a ratio of four to one, thereby enhancing parallelism and reuse efficiency.
- We develop a configurable and highly flexible hardware architecture that supports various neuron modules and features event-driven processing, enabling low-power, always-on epilepsy seizure detection with high-accuracy recognition activated upon event detection.
- A novel non-encoding biomedical circuit architecture based on SNNs is proposed. This architecture leverages the consistency between EEG signals and SNN spiking patterns to implement a non-encoding spiking neural network at the hardware level, thereby further reducing energy consumption and hardware resources while maintaining classification accuracy.

II. Spiking Neural Network model

The bio-plausibility of the SNN, coupled with its spike-based method to EEG signal processing and its low-power consumption characteristics, affords SNN significant advantages in the application of wearable epilepsy seizure detection. This network has no encoding layer and consists of just two layers: locally connected (LC) and classification layers. The pure neural signals are fed into the non-encoding interface, where they undergo event buffering, identifier, and receive additional timestamp information. The concept of the LC layer is borrowed from the deep learning literature [20], [21], [21] and introduced in SNN first introduced by [22]. The neuron model in the LC layer is described in section II-A. The weights of the synapses between them are adjusted using a reconfigurable on-chip learning architecture that employs the triplet-based spike-timing-dependent plasticity (TSTDP) learning rule, as well as the classification layer [23]–[25] discussed in section II-B.

A. Neuron Model

As the basic component of the spiking neural network, the spiking neuron possesses rich computational properties. There exist mainly three spiking neuron models, namely Hodgkin-Huxley model [26], leaky integrate and fire (LIF) model [27] and Izhikevich model [28]. Although the LIF model has the lowest biological plausibility among them, it requires fewer computational operations which makes it popular for hardware implementation. In this work, considering the low computational complexity of the spiking neuron model, we chose the variant versions of the LIF model and the integrate-and-fire (IF) model as the neurons for the SNN [29].

The LIF neuron model includes three stages. All the spikes fired by the presynaptic neurons are picked up and integrated into the membrane potential by the postsynaptic neuron. Only when the membrane potential reaches its membrane threshold, a spike is fired. This event-driven mechanism explains why SNN computation consumes less power. Meanwhile, if the membrane potential of the postsynaptic neuron cannot reach the membrane threshold, it will leak and gradually return to the resting state. The membrane potential $V_i^l(t)$ of neuron i in layer l can be represented as follows:

$$\begin{cases} \frac{\mathrm{d}V_i^l(t)}{\mathrm{dt}} = -\frac{V_i^l(t)-V_{rest}^l}{\tau} + \sum_{j=1}^{n} W_{ji}^l I_j, V_i^l(t) < V_{th}^l(t) \\ V_i^l = V_{rest}^l, V_i^l(t) \geq V_{th}^l \end{cases} \tag{1}$$

where τ denotes the time constant, V_{rest}^{i} denotes the resting membrane potential of layer l , W_{ji} denotes the connection weight between the presynaptic neuron j in layer $l-1$ and the postsynaptic neuron i in layer l, I_{ji} denotes the time-dependent input current from the presynaptic neuron j to the postsynaptic neuron i and V_{th}^{l} denotes the membrane threshold of layer l.

The AI processor for detecting epilepsy seizures continuously detects events using the IF model to conserve energy. To streamline the model and facilitate hardware implementation, we eliminated the leak module and transformed the linear differential model into a discrete one. Typically, we define τ such that Vrest equals zero. Consequently, the discrete model can be formulated as follows:

$$\begin{cases} \tau \frac{\mathrm{d}V_i^l(t)}{\mathrm{d}t} = \sum_{j=1}^{n} W_{ji}^{l} I_j \\ I_i^l = \sum_{j=1}^{n} W_{ji}^{l} S_j^{l-1} \end{cases} \tag{2}$$

B. Learning Rules and Classification Layer

We apply the triplet-based spike timing dependent plasticity (TSTDP) rule, which is proven to be more effective in training SNN than Pair-based STDP (PSTDP), to modify the weights of synapses between the input layer and the locally connected layer [23]. The synaptic TSTDP rules are described as follows:

$$\begin{cases} w(t) \longrightarrow w(t) + A^{+} pre(t) post2(t-\epsilon), if \quad t = t_{post} \\ w(t) \longrightarrow w(t) - A^{-} pre(t) post1, if \quad t = t_{pre} \end{cases} \tag{3}$$

where A^{+} (A^{-}) is the learning rate of post (pre) synaptic activities. ϵ represents a small time offset, while pre, post1, and post2 denote synaptic traces. When the pre-synaptic (post-synaptic) neuron emits a spike, pre (post1, post2) is set to 1, otherwise, it decays exponentially over time.

Traditional SNN usually adopts the winner-take-all mechanism for final classification. In the classification stage, each neuron in the classification layer only votes for one output class assigned to it in the assignment stage. However, it will cause information loss for locally connected SNN because each output neuron is only responsive to a local feature in reality. The vote-for-all mechanism [30], where each neuron in the LC layers is required to vote for all classes upon firing a spike. Considering the hardware-friendly nature of this mechanism and its enhanced classification performance, we apply it to the classification layer. The class with the highest score is regarded as the final predicted classification result and the weights of votes are based on the LC layers' responses to each training sample at the assignment stage.

III. Hardware Implementation

A. Overall Architecture

Fig. 2 shows the overall architecture of the proposed AI processor. It mainly consists of a sensor interface & parameter configuration module, D-A engine, non-encoding interface, VFA classification, neuron cores, shared computation units (SCU), and unified synapse memory (USM). The sensor interface & parameter configuration module is designed for

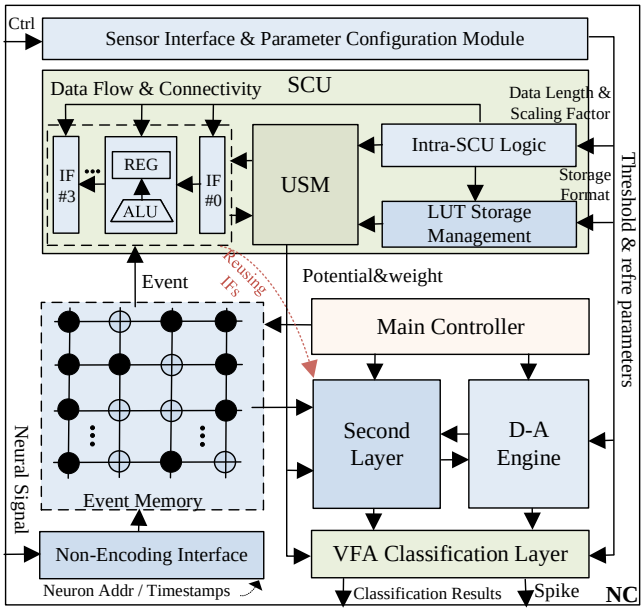


Fig. 2: Overall architecture of the proposed AI processor

interaction and parameter configuration. The main controller is responsible for loading EEG data, managing all system operations, and supplying timestep information. The D-A engine is designed to enable more complex and larger networks by achieving hardware resource integration and reuse. The pure neural signals are fed into the non-encoding interface, where they undergo event buffering identifier, and receive additional timestamp information. The vote-for-all classification, which performs the voting mechanism mentioned in Section II, is implemented in this article using circuitry. The SCU is designed for completing both the SNNs inference computation and the online learning. The USM, situated in close proximity to the SCU, efficiently stores variable-width weights of various networks, connection relationships, and synapse delays. Its strategic placement enhances memory locality, rendering it a crucial component for the implementation of D-ACA. This reconfiguration allows for the switching between two levels of classification (coarse or fine) without the need for standby hardware.

B. Architecture Design Philosophy

1) Sparsity and Architecture: Biomarkers of diseases can be detected or predicted by devices only for a short period [31]. In patients with epilepsy, seizures account for only 0.01% of their lives. This feature can be utilized in seizure detection: the sparsity of abnormal signals. It means that precise identification is necessary only when an abnormal signal occurs so that event-driven processing methods can be implemented. In most cases, the system is in a normal state [2], as illustrated in Fig. 3. Given these characteristics, biomedical AI processors for detecting epilepsy seizures need to have high data processing capacity to achieve accurate analysis within a short period. They also need to have low power consumption to support long-term detection.

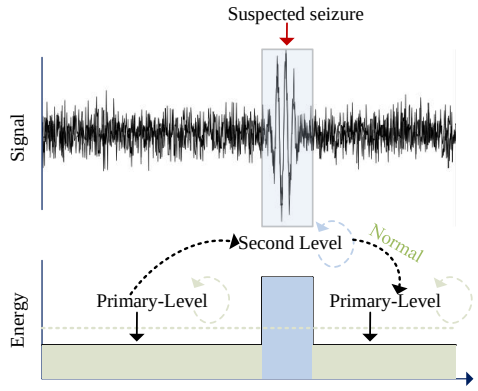


Fig. 3: Sparse processing for abnormal signal detection

Consequently, some researchers attempt to implement a two-level detection method to avoid keeping the precise classifier always on, achieving low power consumption with commendable results [13]. However, it inevitably relies on complex networks, such as ANNs or machine learning, as standby hardware for short-period precise classification. While this method reduces the activation time of the precise classifier and avoids continuous operation, it still results in power and hardware resource wastage. The two-level detection method in [12] employs an event-driven method, utilizing both BNN for primary detection and SNN for precise identification. However, due to the heterogeneity of BNN and SNN, it still requires some standby hardware and cannot achieve hardware fusion across channels.

In this work, we developed an event-driven processing method and a distributed-aggregated classification architecture (D-ACA), which maximizes the utilization of hardware resources on the processor. With this reconfigurable method, we can achieve hardware restructuring across channels, enabling low-power seizure detection with low hardware waste.

2) Design of Architecture: To address the unique characteristics of epilepsy, including short inter-seizure intervals, brief seizure durations, and the need for continuous monitoring around the clock, a novel solution is required in terms of low power and low cost.

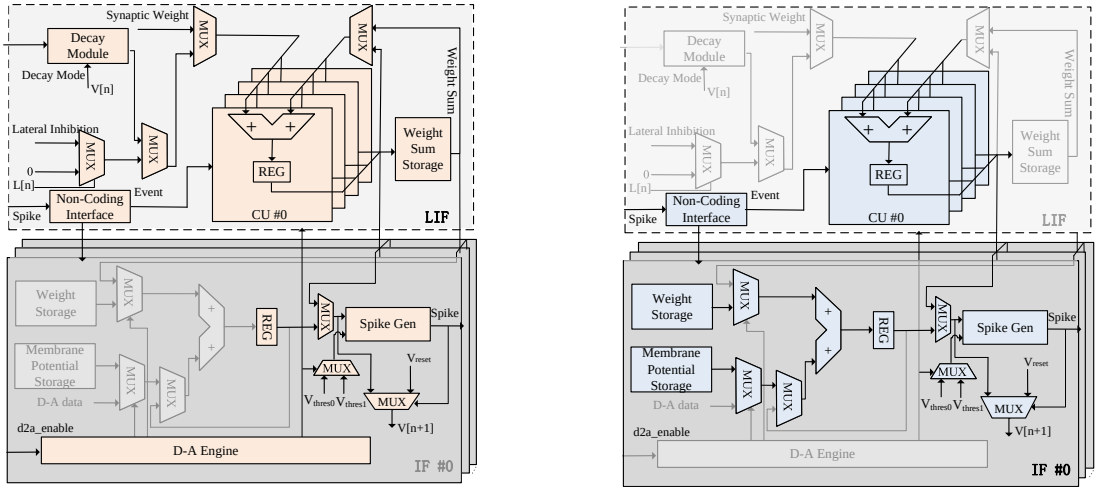
The key issue in the design of the proposed D-ACA for integrating both the primary level and the second level is how to merge the computing, storage, and communication resources required by these two different paradigms with minimal overhead. We utilize the LIF model for second-level and the IF model for primary-level spiking neurons. The primary level outputs only two results: normal and suspected abnormal, while the second level supports the output of final classification results. The behavior of LIF neurons can be abstractly summarized into these processes: membrane potential leakage, weight accumulation, lateral inhibition, and spike generation. IF neurons have two processes: weight accumulation and spike generation [32]. Two-level recognition involves differences not only in the neuron models but also in the synaptic weights (primary level: 4-bit, second level: 16-bit), as well as the calculations for updating neurons. The second level also includes important functions such as delay and lateral inhibition, which are beneficial for improving computational accuracy. Therefore, its scale is relatively large. In fact, SNNs can currently be applied to larger networks and neuromorphic AI processors need to move beyond the mere application for small networks and simple tasks to achieve practical significance. By employing a design that incorporates a non-encoding module and a two-level D-ACA recognition architecture (simple brain and complex brain), we achieve ultra-low power consumption and reduced hardware utilization.

C. D-ACA Paradigms in Circuit Level

The distributed-aggregated classification architecture (D-ACA) is the key to achieving event-driven processing. The AI processor for detecting epilepsy seizures can perform always-on event detection with low power consumption (primary level), and it only switches to the high-accuracy LIF unit SNN-based recognition mode (second level) when events are detected. The primary level monitors the states of 23 channels, marking them as 'suspected abnormal' or normal. If all channels are in the normal state, the processor's cores operate in a distributed manner for energy efficiency. Once any channel is classified as a 'suspected abnormal', the main controller activates the D-A engine and reconfigures all hardware resources to merge them for executing complex algorithms. The always-on event detection across all channels is temporarily paused, allowing entry into the second level for precise identification. This precise identification persists until the final identification result is provided. This reconfigurable method does not affect the accuracy of seizure detection and ensures that the second level optimally utilizes the hardware resources allocated by the primary level across all channels, resulting in significantly improved energy efficiency and hardware utilization.

D-ACA Circuit Schematics: As shown in Fig. 4a and Fig. 4b, the synaptic weights are read from the shared synaptic weight memory and then transmitted to each compute unit (CU). The input spike is broadcast to each CU of the neuromorphic core (NC) in each cycle. To enhance energy efficiency, we configure the hardware to skip computations when the input signal is zero, ensuring that the synaptic weights of each computational unit do not accumulate. Considering that frequent read and write operations on storage can lead to energy waste, when a spike signal is received, the synaptic weights of the CUs are accumulated using membrane potentials from the shared storage. New membrane potentials are then stored in the registers of the current CU. The LIF module, reconstructed by the D-A engine in the second level, is also controlled by the immediate number of instructions configured by the sensor interface & parameter configuration module. The values to be configured include the threshold and leakage values of the current neuron. If the membrane potential of this reconfigured NC exceeds the threshold value, a spike output is generated. Otherwise, a zero value is used as the output.

Due to the configuration of cascaded IF neurons, the computation bit width of the spiking neurons is determined. As previously mentioned, the characteristics of epilepsy allow for cost-effective seizure detection by first identifying anomalies



(a) (b)

Fig. 4: D-ACA Fusion circuit implementation of (a) LIF neuron and (b) IF

and then performing precise recognition. A neuromorphic core consists of four IF neurons (for anomaly detection), which are reconfigured and aggregated into one second-level LIF neuron (for precise recognition). Therefore, the computation bit width for second-level recognition is 16 bits. When the processor operates at the primary level, upon the arrival of a spike, the input spike is simultaneously sent to each connected IF neuron. In contrast to the predominantly sequential updates currently in use [31], [33], we introduce a parallel update strategy for IF neurons, whereby the four IF neurons within the NC can concurrently update four digital neurons in a single operation. To enhance the efficiency of parallel computation, we employ a 16-bit storage width, allowing a single storage address to hold the synaptic weights of four post-synaptic neurons concurrently. Compared to their storage read and write operations, our method achieves a 3× reduction in power consumption. At the second level of operation, each address stores one synaptic weight.

D. Distributed-Aggregated Classification Architecture Flow

Fig. 5 shows the overall operating flow of the D-ACA. The input EEG data is initially received and distributed by the sensor interface & parameter configuration module to the respective non-encoding interfaces to append necessary information for event buffering, and subsequently forwarded to the corresponding neuromorphic core (NC) for classification. The NC operates in always-on SNN mode, extracting neuron membrane voltages based on event information to update corresponding weights and complete neuron updates. When a suspected abnormal signal is detected, the main controller activates the D-A engine to reconfigure the hardware, enabling precise identification through high-precision complex networks. If no suspected abnormal signal is detected, the output is classified as normal. The processor operates in the ultra-low-power SNN inference computation mode for the majority of the time, which is utilized for primary-level classification. To improve the classification accuracy, during the user calibration

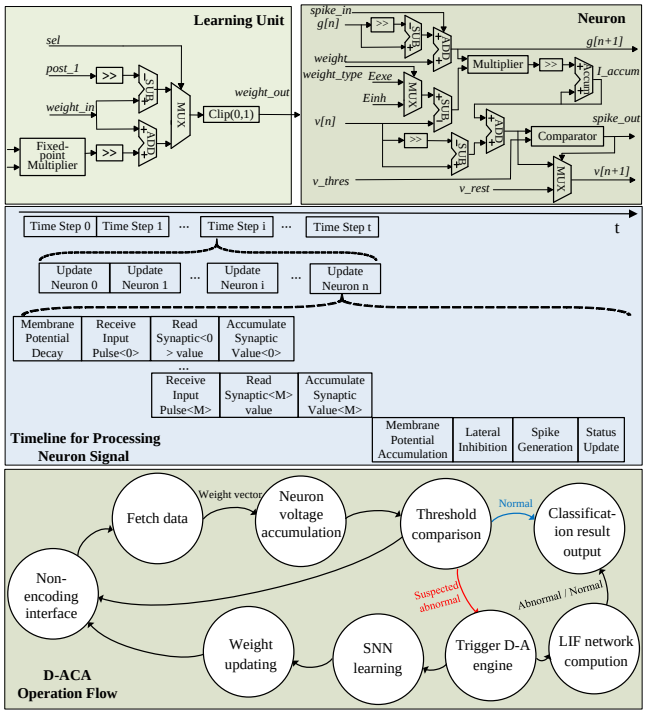


Fig. 5: Hardware circuit & timeline & operation flow of D-ACA

phase, the online learning mode can be enabled to trigger the learning to update the weights. Online learning uses the same hardware as inference but operates at the second level. In practical use, even a small dataset containing user EEG data along with labels (e.g., a 5-minute recording) can significantly enhance accuracy. Moreover, doctors can efficiently label EEG data during routine consultation services with minimal effort.

Neuron: As mentioned above, the neuron cores perform SNN computation. The fetched data is decoded and transmitted to the neurons. Our design supports two modes of neuromorphic computing. In the IF mode, simply accumulating the hardware computation results suffices, which can then be

directly output through g[n+1]. Signal decay is required in the LIF neuromorphic computing mode, which involves the use of multipliers. Subsequently, after synaptic accumulation, a comparator is employed to determine whether to generate a spike, which then outputs the corresponding voltage signals and spikes. For further hardware simplification, the decay of membrane potential and synaptic conductance can also be achieved through the aforementioned bit-shift operations.

Learning Unit: As mentioned in the model described in Section II, synaptic plasticity is characterized as a function of the time difference between sets of three spikes. The TSTDP considers sets of three spikes, i.e., two pre and one post or one pre and two post [34]. For instance, in the case of two pre and one post, the signal is processed through a multiplier, and the resulting output is used for weight update via shift operations and selection signals.

Timeline for Processing Neuron Signal: The timeline for processing one input data. The input data needs to be processed in t time steps. An input data requires processing over t timesteps. At each time step, after the conversion of spikes into events, the states of all neurons in the LC layer must be updated. To complete the update of a neuron, after the decay of membrane potential, the conductance of all the synapses connected to the neuron will be updated to compute the synaptic current accumulated to the membrane potential.

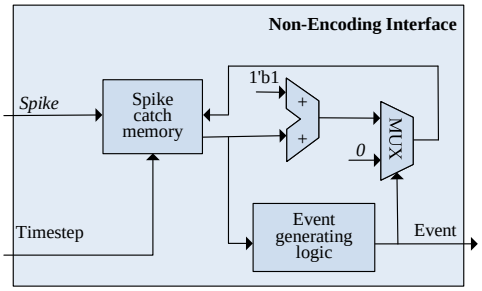


Fig. 6: Hardware implementation of non-encoding interface

E. Non-Encoding Interface

EEG signals differ from other bioelectric signals in that it has a natural spike signal form. Previous works [12], [13], [33] employ SNN architectures based on spike forms for recognizing various bioelectric signals (such as ECG and EMG), which leads to underutilization of the network's compatibility with EEG signals and SNN. Since both neural signals and SNNs are spike-driven, ensuring consistency in signal transmission, we attempt to remove the encoding layer from the network. We also propose a non-encoding module design to reduce power consumption further.

For the processing of neural signals, we are the first to propose and employ a non-encoding design based on the characteristics of our hardware. While omitting an encoding layer can circumvent the complexities of encoding and effectively preserve the spatiotemporal information of neural signals, spikes still necessitate critical information such as delays, connection relationships, and timestamps to operate within the network, as depicted in Fig. 6. Therefore, when spikes are input from an external source, we need to cache them. The non-encoding interface attaches delay (3 bits), connection (1 bit), timestamp information, and caching address to the input spikes, then caches them as events. Subsequently, spikes will operate in the network in the form of events. In the second-level computation, neuron functionality becomes more complex, involving 8-timestep synaptic delays, lateral inhibition, connection relationships, and channel addresses. The main controller uniformly distributes the timestamp information.

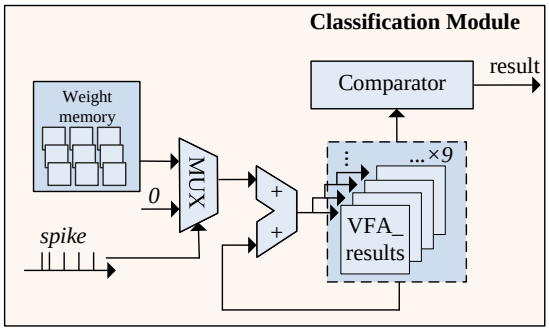


Fig. 7: Hardware of classification layer

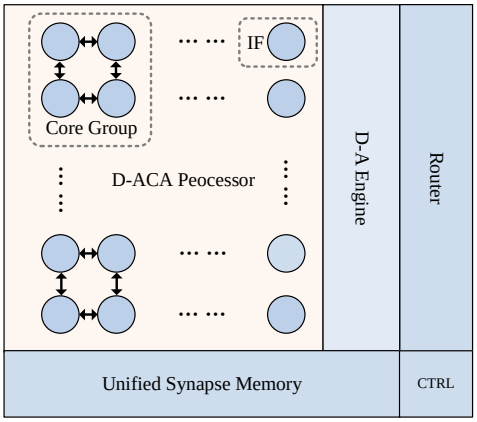


Fig. 8: Inter-core organization of D-ACA processor

F. Classification Layer

As mentioned in Section II, the neurons in the LC layer vote for all classes in the target classification task. The hardware architecture of the vote-for-all (VFA) decoding layer, used in the second-level classifier, is shown in Fig. 7. When the neuron in the LC layer fires a spike, weights between the fired neuron and 9 output neurons are accumulated respectively. The votes accumulated by the output neurons are compared and the class represented by the output neuron with the highest value is the final recognition result.

G. Communication Scheme

Considering the hardware requirements for the extensive parallelism of SNN networks, we further expanded the two-tier architecture by employing a dual-layer communication method to enhance system efficiency and reduce communication latency. In this multi-core communication scheme, we divided

communication into two levels: intra-neuron communication within the core and inter-core communication between cores via on-chip networks, as illustrated in Fig. 8. Firstly, we constructed an IF neuron group within the NC, comprising four IF neurons, enabling data exchange and information transfer through intra-core communication. Secondly, we established an on-chip network within the chip, consisting of multiple cores communicating through on-chip networks, as depicted in Fig. 9. Through this dual-layer communication architecture, we could more effectively handle data transmission and computation tasks in large-scale applications.

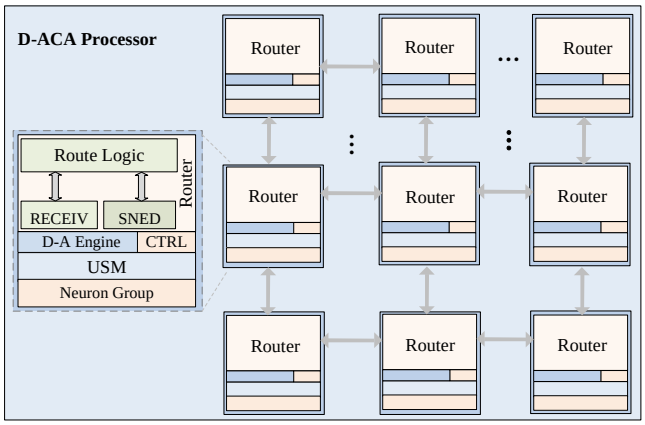


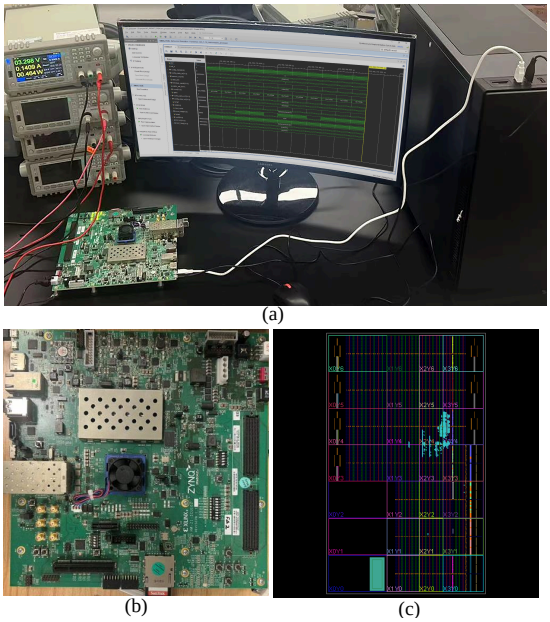
Fig. 9: Multiple cores communication

To achieve efficient data transmission and computation, we introduced routing compute units. These units were responsible for routing data and performing computations, while utilizing send and receive modules for data transmission. Such design not only improved the system's communication efficiency but also reduced communication latency, ensuring reliable support for processing in large-scale applications.

Further, our D-ACA processor supported flexible switching between two levels. Under normal circumstances, the chip operated at the primary level network, with each NC operating in IF mode and supervising signals from one channel. When an abnormal signal was detected in any channel, a signal would be sent, prompting the chip to reconfigure all NCs into LIF mode for precise identification. All hardware resources were dedicated to identifying the abnormal channel signal, ultimately determining if it was indicative of a seizure. Since seizures occur very briefly and infrequently, this operating mode did not result in missed detections or accuracy losses. We could flexibly adjust the system's operating mode according to specific application requirements by configuring the processor to switch between primary and secondary levels. Each NC could only be configured into one mode at any given time, necessitating careful consideration of various factors during configuration to ensure effective utilization of system performance and resources.

IV. EXPERIMENTAL RESULTS

We compare our method against methods such as HNNA [2], CLIC [17], MDMR [1], BioAIP [18], and IoMT [19]. They are also the circuits designed for seizure detection in chips. Since the source codes of some methods are not publicly available, we keep our training/testing setting consistent with them and use the results in their paper.



(a) (b) (c)

Fig. 10: Test system for the proposed AI processor. (a) Overall test system. (b) FPGA board. (c) Utilization of FPGA

A. Experimental Setup

Our proposed design is implemented on the Xilinx ZCU102 FPGA board, as shown in Fig. 10. The clock frequency of our processor is 100MHz, and the number of neurons is set to 4416. The CHB-MIT [35] dataset is used to evaluate the performance of the processor for EEG-based epileptic seizure detection. This dataset categorizes the EEG signals of each subject into s the same subset, totaling 24 subsets, labeled as CHB01-CHB24 [36]–[38]. Each subset corresponds to one electrode, which is also referred to as one channel. In this work, 23 channels are selected based on the international 10–20 system. The EEG signals in the CHB-MIT dataset are continuous in the time domain and are segmented into one-hour segments. A sliding window of 2-seconds length is first used to divide the long-term multichannel EEG signal into epochs. We compare our design with state-of-the-art biomedical AI processors [1], [2], [17]–[19], [39] primarily in terms of energy efficiency, accuracy, hardware resources, and sensitivity (SE). Among these, accuracy represents the ultimate objective, while hardware resources and energy efficiency are the primary contributions of our design. Additionally, hardware resources are crucial for demonstrating low-cost implementation.

We employ a five-fold cross-validation technique to validate the robustness of the proposed D-ACA model. Initially, the EEG dataset is randomly divided into five equal parts, with four parts being utilized as training sets for the model and the remaining one part used as a testing set. Subsequently, the dataset partitioning process is repeated five times, and

the performance of the D-ACA model is comprehensively evaluated using evaluation metrics.

For the patient labeled chb01, there are a total of 40.55 hours of recordings. These recordings are divided into 5 sub-datasets using the k-fold cross-validation method (k = 5), and 5 experiments are conducted to validate the robustness of the model. The final performance is chosen based on the evaluation of the average results. The dataset is categorized into three groups based on the concept of a 2-level architecture, including normal signals (NS), interference signals (IS), and seizure signals (SS). Under the pre-detection of the primary classifier, NS is labeled as normal due to its low power, while IS and SS are forwarded to the secondary classifier for further detection of seizure existence.

In this work, we adopt a patient-specific learning scheme [40]–[42] for the 23 epilepsy patients. This scheme involves the utilization of different model parameters, such as network weights, threshold values, and selected channels, tailored to each patient. The data is divided into training and testing sets at a ratio of 4:1 for each patient, and then the model is trained and tested individually for each patient. This method allows for better capture of the unique features of each patient's EEG signals and enhances the accuracy of seizure detection.

Several evaluation metrics are used to assess the performance of the proposed processor, including accuracy (ACC), sensitivity (SE), specificity (SP), and F1 scores, which can be expressed as follows:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

$$SE = \frac{TP}{TP + FN} \tag{5}$$

$$SP = \frac{TN}{TN + FP} \tag{6}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{7}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. Herein, "positive" and "negative" refer to the "seizure" and "non-seizure," respectively.

B. Performance and Resources

Table I provides a comparative analysis of the D-ACA processor against state-of-the-art biomedical AI processors, including the hybrid neural network processor, heterogeneous computing, two-stage recognition, etc. Due to the relative insufficiency of state-of-the-art FPGA studies for direct comparison, we include several outstanding ASIC studies in our analysis. Although FPGA technologies are at a disadvantage in this context, we still achieve competitive results comparable to those found in ASIC studies. In contrast to traditional architectures, our design enhances the utilization efficiency of hardware resources through cross-paradigm hardware computation integration, without compromising computational performance to achieve hardware reduction. While performing the typical biomedical AI processing tasks of EEG-based seizure detection, the D-ACA processor (including standby power consumption) achieves a low energy per classification (3.73 μJ) and high classification accuracy (98.1%). As shown in Table II, the average sensitivity (SE) is 95.28%, consistent with existing studies [18], [19], indicating that our D-ACA method does not result in additional missed detections while optimizing performance. Benefiting from increased hardware resources and higher data bit width, the IoMT [19] has achieved notable improvements in recognition accuracy. However, the increased hardware resources are disadvantageous for real-time standby and edge detection in the biomedical field. The architecture proposed in this study addresses the excessive use of hardware resources, reducing overall hardware resource consumption by nearly two orders of magnitude while maintaining acceptable recognition accuracy. Incorporating a shared memory design significantly reduces on-chip storage usage, ranking second only to [2], while supporting two-level recognition and achieving a 45.8× reduction compared to [2]. It is suitable for versatile battery-supplied wearable intelligent health detection applications.

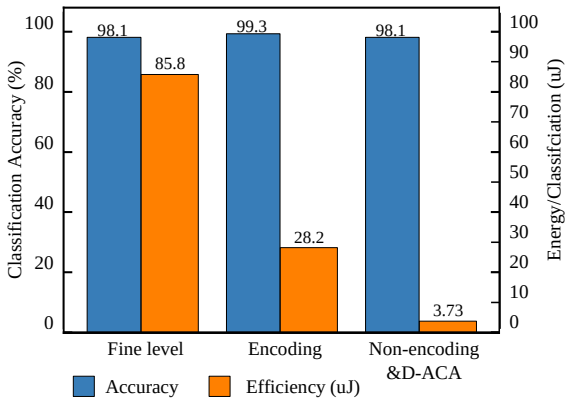


Fig. 11: Classification accuracy & energy consumption

Table II presents the specific experimental results. The experimental outcomes indicate that the ACC and SP of each patient method are 98.1% and 98.35%, signifying a remarkably low false positive rate of the model, which can effectively enhance patient compliance and alleviate their anxiety. Furthermore, an average SE of 95.28% is observed, demonstrating the model's capability in identifying seizure samples (SE represents the ability of the SNN classification model not to miss seizure samples. It is noted that 95.28% of SE is almost identical to the state-of-the-art work [2], [12], [18], [19]). The F1 scores for each patient range between 0.919 and 0.991, with an average of 0.948, further confirming the robust identification ability of the model for seizure samples. In conclusion, the proposed D-ACA model exhibits stability in multiple cross-validation tests and demonstrates significant performance advantages in identifying seizure samples.

Fig. 11 shows the classification accuracy and energy with varying bit widths of the synaptic weights. For the EEG-based seizure detection, the sliding window is 2 seconds at 100 MHz and the average energy consumption per classifi-

TABLE I: COMPARISON OF THIS WORK WITH OTHER PROCESSORS

	HNNA'23 [2]	SVM'20 [39]	CLIC'21 [17]	BioAIP'22 [1]	MDMR'23 [18]	IoMT'23 [19]	This work
Hardware Platform	ASIC	ASIC	ASIC	FPGA	ASIC	FPGA	FPGA
Suport Two-Level	Yes	No	Yes	No	No	No	Yes
SRAM Size	160 kB	35.8 kB	64 kB	N/A	73 kB	441 kB	50 kB
Area	5.06 mm^2	5.54 mm^2	3.51 mm^2	N/A	1.74 mm^2	N/A	N/A
Technology	55 nm	40 nm	180 nm	28 nm	65 nm	28 nm	28 nm
Support Neuron Network	BNN/SNN	No	No	SNN	Fixed NN	No	SNN (D-A)
Data Width	24-bit	12/24-bit	16-bit	16-bit	16-bit	32-bit	4/16-bit
Working Frequency	300 KHz	65 KHz /130 KHz	40 MHz	4 KHz	500kHz/2.5 MHz	100 MHz	100 MHz
Accuracy	99.94%	96.6%	99.6%	97.91%	99.84%	99.9%	98.1%
Dataset	CHB-MIT	CHB-MIT	CHB-MIT	SWEC-ETHZ	EEG:bonn	CHB-MIT	CHB-MIT
Energy Per Classification	1.93 μJ	170.9 μJ	14.2 μJ	490 μJ	2.06 μJ	5.68 μJ	3.73 μJ[a]
Latency(sec)	N/A	0.71	<0.3	6.6	N/A	0.112	0.171

[a] The energy consumption data is derived from the D-ACA architecture, which includes the first-level detection ("standby") stage and the consumption of a 100MHz clock source.

TABLE II: EXPERIMENTAL RESULTS OF SEIZURE DETECTION

Pacient	ACC(%)	SP(%)	SE(%)	F1
chb01	98.15	98.64	93.21	0.948
chb02	98.12	98.63	92.82	0.946
chb03	96.17	98.65	94.86	0.947
chb04	96.10	98.63	93.34	0.949
chb05	98.08	98.64	91.14	0.94
chb06	98.18	95.67	95.15	0.928
chb07	99.16	98.64	98.16	0.923
chb08	98.15	98.65	93.67	0.939
chb09	98.14	98.57	96.13	0.965
chb10	99.09	98.66	99.34	0.991
chb11	98.12	97.68	98.75	0.987
chb12	98.15	98.62	96.11	0.934
chb13	95.14	96.63	93.15	0.941
chb14	98.16	98.65	96.27	0.932
chb15	98.12	98.64	93.42	0.971
chb16	98.17	97.65	96.73	0.951
chb17	98.18	98.65	95.49	0.95
chb18	99.15	98.65	97.48	0.919
chb19	98.11	97.63	92.50	0.931
chb20	98.16	98.65	96.10	0.96
chb21	98.12	100	97.14	0.974
chb22	99.17	98.63	95.24	0.931
chb23	98.19	98.64	95.35	0.945
Average	98.10	98.35	95.28	0.948

cation is 3.73 μJ. We conduct comprehensive experiments, demonstrating that the two-level implementation achieves a 23.7× reduction in energy consumption compared to precise identification alone while maintaining the same recognition rate. Compared with traditional encoding architecture [43], we also achieve significant advantages with 7.56× energy consumption reduction. According to previous studies [1], [17], [18], [39], these methods suffice to fulfill the stringent demands for high accuracy and ultra-low power performance in biomedical applications. As expected, the implementation of the non-coding biomedical circuit architecture achieves optimal energy efficiency with a slight reduction in recognition accuracy, remaining relatively high and surpassing most existing studies. While the recognition rate is marginally lower by 1.2% compared to traditional encoding architectures, there is a significant reduction of 6.73× in power consumption.

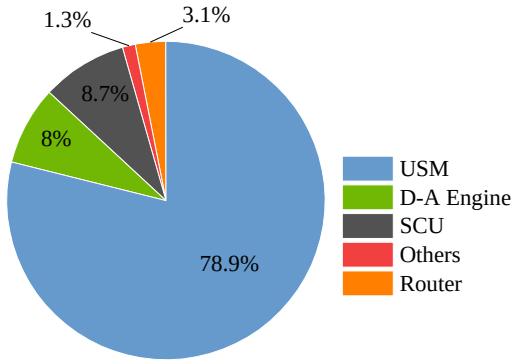


Fig. 12: Hardware resource breakdown

The hardware resource allocation of the D-ACA processor is illustrated in Fig. 12. During biomedical artificial intelligence processing, the USM and SCU occupy 78.9% and 8.7% of the resource, respectively. The USM module is responsible for data storage and caching, while the SCU module executes data analysis and computation tasks. Additionally, the D-A engine and controller of the AI processor occupy 8.0% of the area. The D-A engine serves as the core component of the 2-level

recognition and is responsible for implementing data streaming processing and management, while the controller manages the processor's operational status and scheduling tasks.

TABLE III: RESOURCE UTILIZATION OF SINGLE CORE

Operating Frequency	100MHz	**Total Memory**	2kB
LUTs	1138	**Physical Neuron**	4/1
FFs	1802	**Synapses**	1024

[a] The LUTs of the non-coding interface amount to 209, accounting for 18.3% of the total LUTs.
[b] The FFs of the non-coding interface amount to 276, accounting for 15.4% of the total FFs.

Table III illustrates the resource utilization of a single core. The resources allocated to neuron cores and the VFA decoding layer are predominantly utilized by synapses and adder trees, respectively. Fig. 13 illustrates the reduction in inference complexity achieved by the proposed distributed-aggregated classification architecture, compared to the state-of-the-art existing work in hardware resources [1].

The proposed distributed-aggregated classification architecture achieves a 3.46× reduction in memory usage, while LUTs and FFs experience a reduction of 1.68×. Our minimum data bit-width decreases by 4×, with up to 1024 synapses. Meanwhile, the average accuracy is slightly higher by 0.19%.

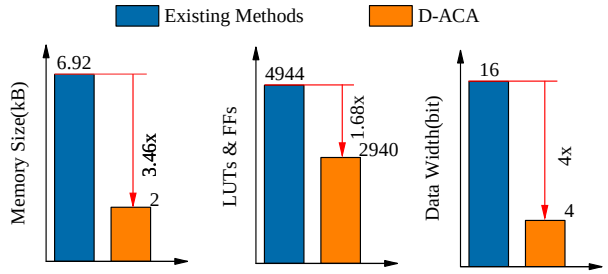


Fig. 13: Comparative analysis of inference complexity reduction between existing methods and distributed-aggregated classification architecture

V. Conclusion

In this work, we design a low-power, low-cost AI processor for epilepsy seizure detection. First, we propose a distributed-aggregated classification architecture (D-ACA), which cross-level reuses computational and storage resources through an event-triggered mechanism. This architecture allows for hardware reuse in a two-stage recognition process, eliminating the need for standby hardware which enhances energy efficiency and reduces hardware resource consumption. In addition, we introduce a novel non-coding biomedical structure based on SNNs. This architecture leverages the consistency between EEG signals and SNN spiking patterns to implement a non-encoding spiking neural network at the hardware level, thereby further reducing energy consumption and hardware resources while maintaining classification accuracy. The processor achieves a low energy per classification (3.73 μJ) and high classification accuracy (98.1%), with at least a 3.46× reduction in memory usage and a 1.68× reduction in hardware resource consumption. The average sensitivity (SE) is 95.28%, consistent with existing work, indicating that our D-ACA method does not lead to additional missed detections.

Acknowledgments

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2023B0303030004, in part by the National Natural Science Foundation of China (NSFC) under Grant 62334014, and in part by the Basic Research Operating Expenses of Universities-Young Teachers Cultivation Programs (No. 24qnpy140).

References

[1] K. F. Razi and A. Schmid, "Epileptic seizure detection with patient-specific feature and channel selection for low-power applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 626–635, 2022.

[2] S. Zhao, J. Yang, J. Wang, C. Fang, T. Liu, S. Zhang, and M. Sawan, "A 0.99-to-4.38 uj/class event-driven hybrid neural network processor for full-spectrum neural signal analyses," *IEEE Transactions on Biomedical Circuits and Systems*, 2023.

[3] K. F. Razi and A. Schmid, "Epileptic seizure detection with patient-specific feature and channel selection for low-power applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 626–635, 2022.

[4] 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.

[5] H. Shan, L. Feng, Y. Zhang, L. Yang, and Z. Zhu, "Compact seizure detection based on spiking neural network and support vector machine for efficient neuromorphic implementation," *Biomedical Signal Processing and Control*, vol. 86, p. 105268, 2023.

[6] Y. Wen, Y. Zhang, L. Wen, H. Cao, G. Ai, M. Gu, P. Wang, and H. Chen, "A 65nm/0.448 mw eeg processor with parallel architecture svm and lifting wavelet transform for high-performance and low-power epilepsy detection," *Computers in Biology and Medicine*, vol. 144, p. 105366, 2022.

[7] Y. Wu, Y. Liu, S. Liu, Q. Yu, T. Chen, and Y. Liu, "Spike-driven gated recurrent neural network processor for electrocardiogram arrhythmias detection realised in 55-nm cmos technology," *Electronics Letters*, vol. 56, no. 23, pp. 1230–1232, 2020.

[8] S. K. Bose, B. Kar, M. Roy, P. K. Gopalakrishnan, L. Zhang, A. Patil, and A. Basu, "Adepos: A novel approximate computing framework for anomaly detection systems and its implementation in 65-nm cmos," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 3, pp. 913–926, 2019.

[9] S.-Y. Hsu, Y. Ho, P.-Y. Chang, C. Su, and C.-Y. Lee, "A 48.6-to-105.2 μw machine learning assisted cardiac sensor soc for mobile healthcare applications," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 801–811, 2014.

[10] S. Stanslaski, J. Herron, T. Chouinard, D. Bourget, B. Isaacson, V. Kremen, E. Opri, W. Drew, B. H. Brinkmann, A. Gunduz *et al.*, "A chronically implantable neural coprocessor for investigating the treatment of neurological disorders," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 6, pp. 1230–1245, 2018.

[11] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "Unpu: A 50.6 tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2018, pp. 218–220.

[12] Y. Wang, H. Luo, Y. Chen, Z. Jiao, Q. Sun, L. Dong, X. Chen, X. Wang, and H. Zhang, "A closed-loop neuromodulation chipset with 2-level classification achieving 1.5-vpp cm interference tolerance, 35-db stimulation artifact rejection in 0.5 ms and 97.8%-sensitivity seizure detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 4, pp. 802–819, 2021.

[13] Y. Wang, Q. Sun, H. Luo, X. Chen, X. Wang, and H. Zhang, "A closed-loop neuromodulation chipset with 2-level classification achieving 1.5-vpp cm interference tolerance, 35-db stimulation artifact rejection in 0.5 ms and 97.8%-sensitivity seizure detection," in *2020 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2020, pp. 404+.

[14] Y.-Y. Hsieh, Y.-C. Lin, and C.-H. Yang, "A 96.2-nj/class neural signal processor with adaptable intelligence for seizure prediction," *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 167–176, 2022.

[15] Y. Zhao, Z. Shang, and Y. Lian, "A 13.34 μw event-driven patient-specific ann cardiac arrhythmia classifier for wearable ecg sensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 2, pp. 186–197, 2019.

[16] Z. Taufique, B. Zhu, G. Coppola, M. Shoaran, and M. A. B. Altaf, "A low power multi-class migraine detection processor based on somatosensory evoked potentials," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 5, pp. 1720–1724, 2021.

[17] Y. Wang, H. Luo, Y. Chen, Z. Jiao, Q. Sun, L. Dong, X. Chen, X. Wang, and H. Zhang, "A closed-loop neuromodulation chipset with 2-level classification achieving 1.5-vpp cm interference tolerance, 35-db stimulation artifact rejection in 0.5 ms and 97.8%-sensitivity seizure detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 4, pp. 802–819, 2021.

[18] J. Liu, J. Fan, Z. Zhong, H. Qiu, J. Xiao, Y. Zhou, Z. Zhu, G. Dai, N. Wang, Q. Liu *et al.*, "An ultra-low power reconfigurable biomedical ai processor with adaptive learning for versatile wearable intelligent health monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, 2023.

[19] W. Zhao, Y. Wang, X. Sun, S. Zhang, and X. Li, "Iomt-based seizure detection system leveraging edge machine learning," *IEEE Sensors Journal*, 2023.

[20] Y.-h. Chen, I. L. Moreno, T. Sainath, M. Visontai, R. Alvarez, and C. Parada, "Locally-connected and convolutional neural networks for small footprint speaker recognition," 2015.

[21] A. Bakiya, K. Kamalanand, V. Rajinikanth, R. S. Nayak, and S. Kadry, "Deep neural network assisted diagnosis of time-frequency transformed electromyograms," *Multimedia Tools and Applications*, vol. 79, no. 15, pp. 11 051–11 067, 2020.

[22] D. J. Saunders, D. Patel, H. Hazan, H. T. Siegelmann, and R. Kozma, "Locally connected spiking neural networks for unsupervised feature learning," *Neural Networks*, vol. 119, pp. 332–340, 2019.

[23] D. Ma, J. Shen, Z. Gu, M. Zhang, X. Zhu, X. Xu, Q. Xu, Y. Shen, and G. Pan, "Darwin: A neuromorphic hardware co-processor based on spiking neural networks," *Journal of Systems Architecture*, vol. 77, pp. 43–51, 2017.

[24] S. Wang, W. A. Chaovalitwongse, and S. Wong, "Online seizure prediction using an adaptive learning approach," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 12, pp. 2854–2866, 2013.

[25] J.-E. Le Douget, A. Fouad, M. M. Filali, J. Pyrzowski, and M. Le Van Quyen, "Surface and intracranial eeg spike detection based on discrete wavelet decomposition and random forest classification," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 475–478.

[26] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical journal*, vol. 1, no. 6, pp. 445–466, 1961.

[27] J. Vreeken *et al.*, "Spiking neural networks, an introduction," 2003.

[28] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on neural networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[29] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[30] M. Meng, X. Yang, L. Bi, J. Kim, S. Xiao, and Z. Yu, "High-parallelism inception-like spiking neural networks for unsupervised feature learning," *Neurocomputing*, vol. 441, pp. 92–104, 2021.

[31] Y. Liu, Z. Wang, W. He, L. Shen, Y. Zhang, P. Chen, M. Wu, H. Zhang, P. Zhou, J. Liu *et al.*, "An 82nw 0.53 pj/sop clock-free spiking neural network with 40μs latency for alot wake-up functions using ultimate-event-driven bionic architecture and computing-in-memory technique," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 372–374.

[32] G. K. Chen, R. Kumar, H. E. Sumbul, P. C. Knag, and R. K. Krishnamurthy, "A 4096-neuron 1m-synapse 3.8-pj/sop spiking neural network with on-chip stdp learning and sparse weights in 10-nm finfet cmos," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 992–1002, 2018.

[33] R. Mao, S. Li, Z. Zhang, Z. Xia, J. Xiao, Z. Zhu, J. Liu, W. Shan, L. Chang, and J. Zhou, "An ultra-energy-efficient and high accuracy ecg classification processor with snn inference assisted by on-chip ann learning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 832–841, 2022.

[34] H. Zheng and Y. Yi, "Enhancing snn training performance: A mixed-signal triplet reconfigurable stdp circuit with multiplexing encoding," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2023, pp. 1–5.

[35] P. PhysioBank, "Physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[36] A. R. Ozcan and S. Erturk, "Seizure prediction in scalp eeg using 3d convolutional neural networks with an image-based approach," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 11, pp. 2284–2293, 2019.

[37] Y. Zhang, S. Yao, R. Yang, X. Liu, W. Qiu, L. Han, W. Zhou, and W. Shang, "Epileptic seizure detection based on bidirectional gated recurrent unit network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 135–145, 2022.

[38] P. Boonyakitanont, A. Lek-Uthai, and J. Songsiri, "Scorenet: a neural network-based post-processing model for identifying epileptic seizure onset and offset in eegs," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2474–2483, 2021.

[39] 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.

[40] X. Xu, Y. Zhang, R. Zhang, and T. Xu, "Patient-specific method for predicting epileptic seizures based on drsn-gru," *Biomedical Signal Processing and Control*, vol. 81, p. 104449, 2023.

[41] M. Kaleem, A. Guergachi, and S. Krishnan, "Patient-specific seizure detection in long-term eeg using wavelet decomposition," *Biomedical Signal Processing and Control*, vol. 46, pp. 157–165, 2018.

[42] C. Cheng, B. You, Y. Liu, and Y. Dai, "Patient-specific method of sleep electroencephalography using wavelet packet transform and bi-lstm for epileptic seizure prediction," *Biomedical Signal Processing and Control*, vol. 70, p. 102963, 2021.

[43] E. Donati, M. Payvand, N. Risi, R. Krause, and G. Indiveri, "Discrimination of emg signals using a neuromorphic implementation of a spiking neural network," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 795–803, 2019.

Qiang Zhang the School of Microelectronics Science and Technology, Sun Yat-sen University, Zhuhai 519000, China where he is a Ph.D. candidate. His current research interests include ultra-low-power biomedical AI processors, EEG signal perception, domain-specific architecture for artificial intelligence, and neuromorphic computing.

Mingyue Cui received a B.Sc. degree in Software Engineering from Chongqing Normal University in 2014, and received an M.Sc. degree in Software Engineering, and a Ph.D. degree in Computer Science from Sun Yat-sen University, Guangzhou, Guangdong, China, in 2017, and 2022, respectively. He was a Visiting Student at the Technical University of Munich, Germany (November 2021–November 2022). He is currently working as a joint Postdoctoral Fellow with the School of Computer Science and Engineering at Sun Yat-Sen University, China. His research interests are in the area of 3D vision and entropy coding.

Yue Liu received the B.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2022. He is currently pursuing an M.S. degree at the School of Sun Yat-sen University, His current research interests include digital integrated circuit design of Spiking Neural Networks and hardware acceleration.

Weichong Chen received the B.Sc. degree in microelectronics from the Guangdong University of Technology, Guangzhou, China, in 2019, and the M.Eng. degree in integrated circuits engineering from the School of Electronics and Information Technology, Sun Yat-sen University (SYSU), in 2021. He is currently pursuing a Ph.D. degree in electronic science and technology with the School of Microelectronics Science and Technology, Sun Yat-sen University (SYSU), Zhuhai, China. His research mainly focuses on low power circuits based on Non-volatile devices and neural network computing

Zhiyi Yu (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Fudan University, Shanghai, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California at Davis, Davis, CA, USA, in 2007.

He was with IntellaSys Corporation, Cupertino, CA, USA, from 2007 to 2008. He was an Associate Professor with the Department of Microelectronics, Fudan University, from 2009 to 2015. He is currently a Professor with the School of Microelectronics Science and Technology, Sun Yat-sen University, Zhuhai, China. He has authored four books (chapters) and over 140 articles, and has granted more than 20 patents. His current research interests include digital VLSI design and computer architecture, including in-memory computing, neuromorphic computing, and many-core processors.

Dr. Yu serves as the TPC Chair/Co-Chair for IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA) and Asia-Pacific Signal and Information Processing Association (APSIPA); a TPC Member for many conference committees, such as Asian Solid-State Circuits Conference (ASSCC) and IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SOC); and an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS.