Seizure-Cluster-Inception CNN (SciCNN): A Patient-Independent Epilepsy Tracking SoC With 0-Shot-Retraining

Chne-Wuen Tsai, *Member, IEEE*, Rucheng Jiang, *Member, IEEE*, Lian Zhang, *Student Member, IEEE*, Miaolin Zhang, *Student Member, IEEE*, and Jerald Yoo, *Senior Member, IEEE*

***Abstract*—Epilepsy tracking System-on-Chips (SoC) usually perform patient-specific classification to deal with the patient-to-patient seizure pattern variation from a surface electroencephalogram (EEG). However, the patient-specific classifier training requires the EEG signals from the target patients a priori, which involves costly and time-consuming hospitalization for the inpatient data collection. To address this issue, this paper presents a patient-independent epilepsy tracking SoC that is trained with pre-existing databases and can be directly deployed to the target patients without collecting their data and performing cumbersome patient-specific training beforehand. The proposed SoC adopts a Seizure-Cluster-Inception Convolutional Neural Network (SciCNN) Neural Processor (SNP) to reduce SRAM access rate by 179.05× with the Kernel-Wise Pipeline (KWP). The 22-Ch. SoC achieves event-based sensitivity of 90.3%/90.4%/83.3% and specificity of 93.6%/95.7%/88.6% on unseen patients from CHB-MIT database/EU database/local hospital patient, respectively.**

***Index Terms*—Cross-patient, epilepsy management, hardware, inter-patient variation, intracranial EEG (iEEG), kernel wise pipeline (KWP), neural network, neural processor, non-patient-specific, patient-nonspecific, patient-to-patient variation, seizure classification, seizure detection, surface EEG.**

I. Introduction

PATIENT-SPECIFIC classifiers are widely adopted by the epilepsy tracking SoC to perform timely and accurate electrical stimulation treatment upon seizure onset to suppress the seizure. Seizure, induced by epilepsy, is a life-threatening neurological disorder that is suffered by more than 50 million population world-wide [1]. Electroencephalogram (EEG) is the recording of the electrical activity in the brain that is widely used for seizure detection. An epilepsy tracking SoC usually senses the EEG of the target patient with an Analog Front-end (AFE), which amplifies and digitalizes the EEG signals; and a Digital Back-end (DBE) subsequently processes and classifies the digitalized EEG into either seizure class or normal class [2]. Furthermore, these seizure tracking SoC needs to be low energy consuming for monitoring over an extended period powered by a small battery, energy harvesting or body-coupled powering [3].

A challenging aspect of detecting seizure/epilepsy from surface EEG is, a pattern may be a seizure for patient A, while the same pattern can be a normal pattern for patent B (inter-patient variations) [2]; Due to the inter-patient variation manifested in the seizure patterns of different patients, patient-specific classification is usually applied to learn and classify the specific seizure pattern of the target patient. EEG signals of both epileptic periods and normal periods are first recorded during hospitalization. After sufficient seizure events have been collected, patient-specific classifiers are trained to differentiate the epileptic EEG patterns from the normal EEG patterns (Fig. 1). Surface EEG is a popular and simple-to-obtain bio-signal for seizure detection. Several works have applied machine learning to learn the patterns of the surface EEG for the seizure monitoring. Two linear Support Vector Machines (SVMs) are used to analyze the spectral domain frequency of the surface EEG for detecting the patient-specific seizure based on hysteresis voting, achieving sensitivity of 95.7% and specificity of 98% [4]. To reduce the hardware-intensive classification, a patient-specific event-triggered 2-level classifiers are used; where a complexed and more powerful classifier will only be triggered when a classifier that is simpler, yet biased to seizure detection, detects a seizure [5]. It reduces the complexed classification triggering rate by 7×, consequently reducing the power consumption. Since the traditional classifiers require manual feature engineering that could lead to inefficiency while deciding the suitable features for the classification, neural network-based seizure-detection classifiers have emerged due to their ability of automatic feature mining. Convolutional Neural Network (CNN) is implemented in the patient-specific SoCs and achieve accuracy of 99.8% [6], sensitivity/specificity of 92.2/95.1% [7], and sensitivity/F1 score of 99.95%/0.93 [8], respectively. However, these three works only test with limited

Manuscript received 4 July 2023; revised 31 August 2023; accepted 1 October 2023. Date of publication 25 October 2023; date of current version 9 January 2024. This work was supported by National Research Foundation, Singapore under grant NRF-000214-00. This paper was recommended by Associate Editor R. Genov. *(Corresponding author: Jerald Yoo.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the National University of Singapore's Institutional Review Board (IRB)'s protocol approval number NUS-IRB-2021-1008.

Chne-Wuen Tsai and Jerald Yoo are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117585, and also with The N.1 Institute for Health, Singapore 117456 (e-mail: tsaicw@u.nus.edu; jyoo@nus.edu.sg).

Rucheng Jiang, Lian Zhang, and Miaolin Zhang are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117585 (e-mail: rucheng.jiang@u.nus.edu; lian.zhang@u.nus.edu; miaolin.z@u.nus.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TBCAS.2023.3327509.

Digital Object Identifier 10.1109/TBCAS.2023.3327509

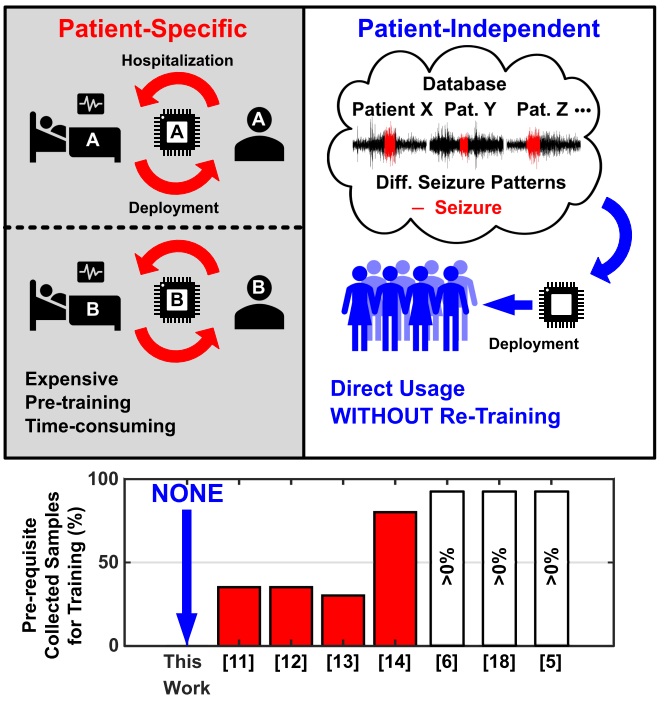

Fig. 1. Comparison between the training/deployment steps of a patient-specific and a patient-independent epilepsy tracking SoC.

dataset and lack of thorough verification. To obtain the EEG signal with less noise, intracranial EEG (iEEG) is also used as the bio-signal for the seizure monitoring. Compared to surface EEG, iEEG contains cleaner signal with less noise due to the direct contact with the brain without the interference of the skull and the scalp, which deteriorates the spectral energy in high frequency bands (64–500 Hz). Based on the iEEG signal of the target patients, spectral energy, phase locking value, and cross-frequency coupling are extracted as the features to detect the patient-specific seizure events, achieving event-based sensitivity of 97.7% and false detection rate of 0.185/hr [9]. However, utilizing iEEG comes with the trade-off of invasive surgery, which is less welcoming due to the surgery and the potential side effects. Nevertheless, regardless of the classifier types (traditional classifier/deep learning) and EEG recording types (surface/intracranial), these patient-specific classifiers require patients to be hospitalized for the inpatient EEG signal recording and the follow-up analysis by the doctors to identify the seizure patterns and the seizure occurrence regions of the brain. As 30% of the epileptic patients come from a low- or middle-income background [1], this treatment could be almost unaffordable, as the seizure occurrence interval could be as long as 97 hours [10].

To avoid prolonged hospitalization duration, online learning/tuning is an emerging technique presented in recent works. Support Vector Machine (SVM) is a widely used patient-specific classifier to detect seizure accurately. On-chip SVM retraining was achieved by using techniques such as $L_{0/1}$ alternative direction method of multipliers (ADMM), pointer-based matrix multiplication, and rearranging the calculation sequence of matrix inversion. These techniques successfully reduce the SVM retraining power consumption and the latency to a hardware-friendly level [11]. However, a big pool of data points needs to be collected and to be stored on-chip before performing the retraining. To further decrease the power consumption and the latency of updating the classifier, SVM tuning is achieved by editing the support vectors pool, which contains the support vectors that have been trained off-chip to form the classification boundary [10], [12]. Compared to entirely retraining a new SVM classifier, tuning the SVM approximates the retraining effect while requiring $>800\times$ less clock cycles. However, it needs supervised input to label the data during the online tuning. To eliminate the manual labeling, a logistic regression (LR) classifier is implemented with unsupervised online learning [13]. Based on the stochastic gradient descent and a series of confidence threshold comparisons, the weights of the LR classifier could be updated in an unsupervised manner. Nevertheless, prior inpatient EEG data collection of the target patient is still unavoidable for all the aforementioned online learning/tuning methods to firstly build a patient-specific classifier.

Aside from the long seizure occurrence interval, the seizure occurrence brain region is also patient dependent. Seizure can be categorized into generalized seizure and focal seizure. The former has the seizure pattern that is observable from the entire brain, whilst the latter only shows the seizure pattern in certain regions. To cover more possible epileptic regions for higher detection precision, a 256-channel seizure detection SoC is presented [14]. With hardware sharing techniques and selectable sub-channels, it achieves 1.51μW/channel power consumption, event-based sensitivity of 95.6%/94.0%, and specificity of 96.8%/96.9% on surface EEG/iEEG databases. However, the channels selection and the classification still require patient-specific training in advance.

To fully eliminate the inpatient data collection and the data analysis of the target patient before the actual deployment, this paper presents a 0-shot-retraining patient-independent epilepsy tracking SoC that implements Seizure-Cluster-Inception Convolutional Neural Network (SciCNN) as the classifier [15]. Patients do not need to be hospitalized for data collection nor data analysis. Instead, they only need to undergo a 2-min automatic on-chip calibration for the classifier adaptation to address the patient-to-patient variation. This significantly increases the practicality of the seizure detection SoC for those patients who cannot afford the expensive incurred cost due to the inpatient data collection. Besides, to capture the seizure events from all the possible regions on different patients with the focal-seizure type, our SoC comprises 22 channels of EEG recording channels (all the channels used in CHB-MIT database, based on 10-20 electrode placement system). Furthermore, both the calibration and the classification of the EEG signal can be performed without any manual input for the labeling and the training. To decrease power consumption while adopting the inception-based SciCNN on a resource-constrained SoC, we implemented three power-efficient hardware techniques, namely Sensor-Stationary Process Elements (SS-PEs), Ping-Pong Layer Buffer (PP-LB), and Kernel-Wise Pipeline (KWP).

This paper is organized as follows: Section II presents the system architecture and motivation. Section III describes the

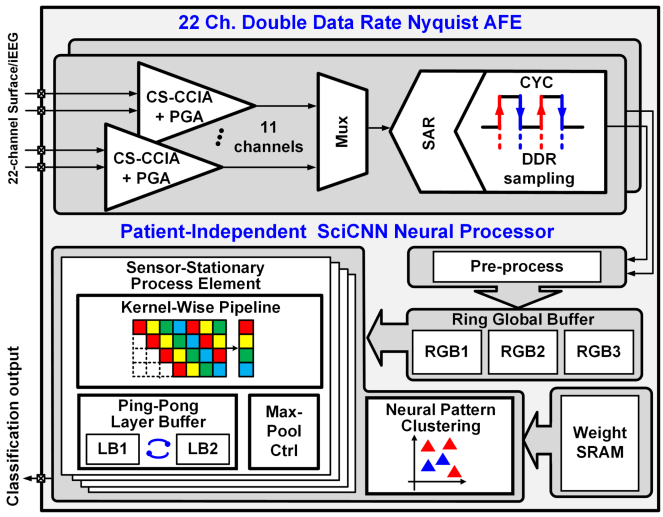

Fig. 2. System architecture.

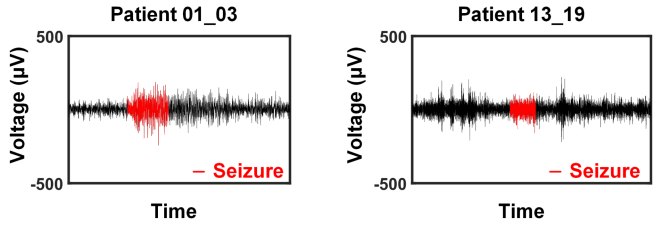

Fig. 3. EEG recordings from two different CHB-MIT patients (patient 01 and patient 13).

design considerations and circuit novelty of SNP. Section IV evaluates and discusses the measurement results. Section V concludes the paper.

II. System Architecture

Fig. 2 shows the proposed patient-independent epilepsy tracking SoC. 22-Ch. differential EEG is used to cover the possible seizure origins of different unseen patients. Chopper Stabilized-Capacitive Coupled Instrumentation Amplifiers (CS-CCIAs) senses and amplifies the EEG signals with low noise and small footprint [12]. To reduce the footprint and avoid the device mismatch, time-sharing Double Data Rate Nyquist Analog Front-End (DDR-NQAFE) is implemented [16]. 11 CS-CCIAs are time-shared by one Double-Data-Rate 2-Stage Pipelined ADC, where the amplified signal of each amplifier is digitalized in a Time-Division Multiplexing (TDM) manner. Then, Seizure-Cluster-Inception Convolutional Neural Network (SciCNN) Neural Processor (SNP) pre-processes (normalization) and stores the digitalized data with three Ring Global Buffers (RGBs), which achieve classification of 2s time window with 1s overlapping. Sensor-Stationary Process Elements (SS-PEs) subsequently perform SciCNN processing with the data stored in two out of the three RGBs, while the remaining one RGB receives the incoming data. Thanks to the 1-D filter used by all the layers in SciCNN, the data from different sensors can be processed independently by all four SS-PEs. This parallel processing enhances the scalability and reduces the overall system clock rate by $\sim 4\times$. Within each SS-PE, the hardware resources are reused for both the initial bandpass filtering and the MAC processes in the forward propagation of SciCNN. During which, two Ping-Pong Layer Buffers (PP-LBs) are implemented to reduce the SRAM size for saving the intermediate data by only storing the output of the previous layer and the current layer; Kernel-Wise Pipeline (KWP) further reduces the power consumption by reducing the data SRAM access rate, which is achieved by computing all the partial sums (psum) that involve the current data point before reading the next data from data SRAM. Finally, Neural Pattern Clustering (NPC) classifies the input data into either seizure data or normal data, which is achieved by checking the label of the cluster centroid that lies the closest to the feature point. The labels are assigned to all 256 cluster centroids during the 1-time 2-min on-chip calibration.

III. SciCNN Neural Processor

Seizure-Cluster-Inception Convolutional Neural Network Neural Processor (SNP) is a hardware accelerator that realizes Seizure-Cluster-Inception Convolutional Neural Network (SciCNN) for the seizure classification. As the computation complexity of the neural network is normally higher than the traditional classifiers, Sensor-Stationary Process Elements (SS-PEs), Ping-Pong Layer Buffer (PP-LB), and Kernel-Wise Pipeline (KWP) are adopted to reduce the power consumption by reducing the sampling clock, SRAM size, and SRAM access rate, respectively.

A. Patient-Independent Classification

To perform a good classification, choices of feature selection and the classifier type are crucial. Traditional classifiers require manual feature engineering to decide the features for performing the data classification. Researchers have suggested many feature types for the seizure detection application, such as spectral energy, phase-locking value, cross-frequency coupling, and line length. Regarding the classifiers, Linear Support Vector Machine (L-SVM) [4], Non-Linear Support Vector Machine (NL-SVM) [11], [12], [17], Decision Tree [18], [19], [20] and Logistic Regression [13] are shown to perform well in the patient-specific classification. However, these techniques would face difficulties in the *patient-independent* classification due to the inter-patient variation. As shown in Fig. 3, two patients from CHB-MIT database show different EEG signal characteristics during the seizure event. This causes confusion to the classifier, especially if the selected features cannot identify the differences. In contrast, deep learning has the advantage of mining the features automatically, which excels in image classification, facial recognition, and natural language processing [21]. Vanilla CNN has been adopted by patient-specific seizure-detection devices [6], [7], [8]. However, vanilla CNN also struggles to perform patient-independent seizure detection well (Fig. 4).

Fig. 4 shows the simulation results of different types of the aforementioned classifiers. Seven bandpass frequency bands (0–28 Hz with 4 Hz bandwidth) are used as the features of the traditional classifiers; 22 channels $\times$ 128 samples $\times$ [1 raw + 7 bandpass frequencies ranging 0–28 Hz with 4 Hz bandwidth] is resized into 227 $\times$ 227 $\times$ 3 and 224 $\times$ 224 $\times$ 3 to fit the

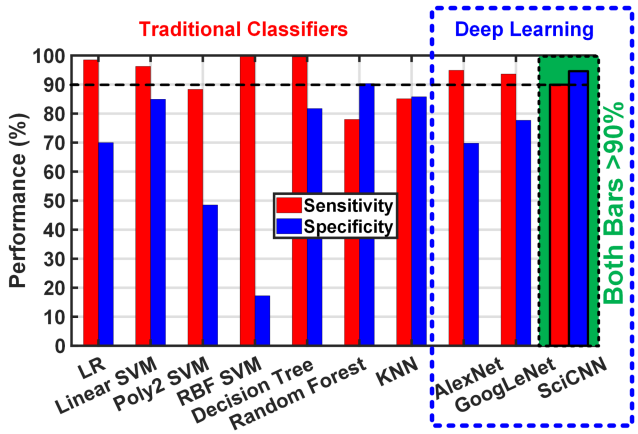

Fig. 4. Simulation of the patient-independent classification performance performed by both traditional classifiers (154-weight LR; 8096-SV linear SVM; 133-SV poly2 SVM; 27884-SV RBF SVM; 1519-node DT; 1024-tree random forest; 256-centroid KNN) and deep learning classifiers; these classifiers are constructed in-house and being fed with spectral features of 8 frequency bands.

input size of AlexNet [22] and GoogLeNet [23], respectively. AlexNet is one of the earliest successful vanilla CNNs that achieves satisfactory results in image classification. Whereas GoogLeNet is an improved CNN that incorporates inception modules in each layer. Multiple inception modules utilize their different filter sizes to extract different features from the same source, hence achieving higher accuracy [23]. Furthermore, the neural networks in this simulation adopt transfer learning by initializing with pre-trained weights in the layers and replacing the last fully connected layers with the proper sizes, which helps in achieving higher accuracy in the application where the sample size is small [24]. More recent efficient networks such as ResNet [25] and MobileNet [26] are not included in the simulation as they will introduce computation and storage overhead due to the shortcut connection (residual blocks). Fig. 4 shows event-based sensitivity and specificity as the verification metrics of the classifiers to detect seizure and normal EEG signal, respectively. These two metrics are crucial to justify the classification ability for not resulting in too many misdetections and false alarms. As the result, all the traditional classifiers, vanilla CNN (AlexNet) and inception-based CNN (GoogLeNet) fail to achieve >90% event-based sensitivity and >90% specificity simultaneously. Hence, we present Seizure-Cluster-Inception Convolutional Neural Network (SciCNN), a 0-shot-retraining patient-independent seizure-detection classifier, which can achieve >90% event-based sensitivity and >90% specificity simultaneously.

B. Seizure-Cluster-Inception CNN (SciCNN)

SciCNN comprises two parts: a 3-layer inception-based CNN (iCNN) and a Neural Pattern Clustering (NPC) layer. iCNN functions as a feature extraction engine and NPC layer functions as a classifier. Fig. 5 shows the convolution structure of one layer in iCNN. Compared to the image classification performed by the general-purpose CNN classifiers, the 'images' of the seizure signals used in this work have a different structure.

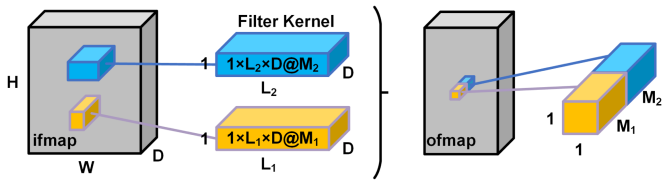

Fig. 5. Convolution structure of inception-based convolutional neural network (iCNN).

In the seizure signal images, height H represents the signals detected by different sensors; width W shows the data samples in a time window; depth D contains the filtered signals with different filtering frequencies. Due to the geometric location of the sensors, neighboring rows might not have strong correlation. Therefore, 1D filter kernels are used in all the SciCNN layers. In contrast to vanilla CNN, inception-based CNN can extract the features from the same source with different filter kernel sizes, hence resulting in higher accuracy [23]. As shown in Fig. 5, two sizes of the filters with different lengths L (shown in different colors) perform convolution with the same input feature map (ifmap). The yielded output feature map (ofmap) is subsequently formed by concatenating the ofmaps generated by all (M) filter kernels. The overall SciCNN structure is shown in Fig. 6. iCNN first receives the digitalized EEG signal and structures it into a 3D image. As the spectral features show promising results in seizure detection, seven filters have been applied to the normalized raw signal to obtain the extra depth dimensions (D) of the iCNN input. This forms an ifmap of Top Layer with the dimensions of H = 22 sensor channels, W = 128 data samples, and D = 1 raw signal + 7 frequency bands. Then, Middle Layer and Deep Layer extract the patient-independent features of the EEG signals and form a 16-D feature point in NPC Layer. Neural Pattern Clustering (NPC) subsequently performs the classification while addressing the inter-patient variation.

Taking the inspiration from k-means clustering, NPC depicts the clusters distribution of the EEG patterns on a patient-independent feature space, which is the feature space that the feature point generated by iCNN is projected on. Fig. 7 shows the training process done in MATLAB 2020b with the solver of stochastic gradient descent with momentum (SGDM). Momentum value γ is set at 0.9, L_2 is set at 10^{-6}, initial learning rate is set at 10^{-5}, which decreases by half every 20 epochs, while the maximum number of epoch is set at 40. First, 256 NPC centroids are randomly initialized on the feature space. Then, a mini-batch of 50 samples forward propagate through iCNN to generate 50 feature points. All the feature points in one mini-batch are deliberately chosen to have the same labels (seizure/non-seizure) and from the same patient to avoid the confusion caused by inter-patient variation, where the same pattern could have different labels on different patients. The Euclidean distance between the cluster centroids and the centers of the feature points that lie the closest to them is used as the loss function to be minimized. As the results, the cluster centroids move towards the centers of the feature points. This increases the ability of the clusters to reflect the EEG pattern more effectively. Meanwhile, the weights of iCNN will be updated based on Euclidean distance between the

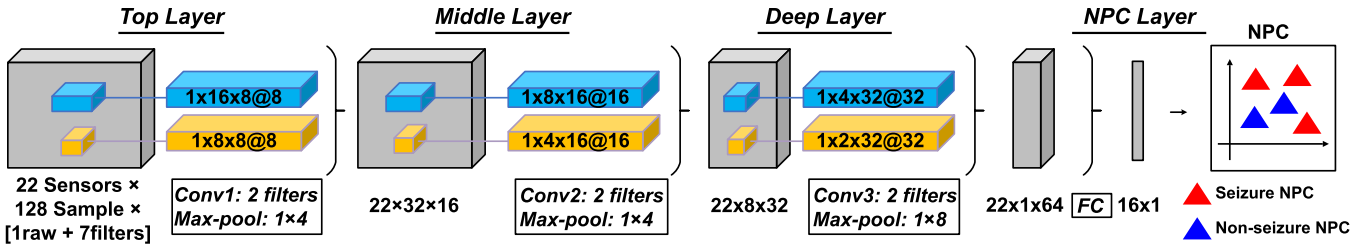

Fig. 6. SciCNN detailed structure.

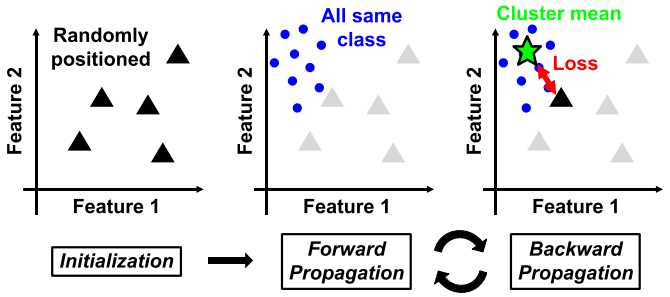

Fig. 7. Training of Neural Pattern Clustering (the number of feature dimension, NPC centroids, and feature points in one mini-batch have been reduced for the ease of visualization).

feature points and the cluster centroids as well. Furthermore, as different patients in the training group have different number of seizure samples, a factor $\gamma_{patient}$ is added to balance the significance of the update of the weights in each mini-batch for alleviating overfitting (1). $\gamma_{patient}$ is inversely proportional to the number of the seizure samples of one patient $N_{patient}$. $\lambda_{patient}$ is a hyperparameter, setting at 0.8, to alter the effect of $\gamma_{patient}$. N is the array that contains seizure sample counts of each patient in the training group. $\delta_{balanced}$ is the new change of weight after taking $\gamma_{patient}$ into account (2).

$$\gamma_{patient} = \lambda_{patient} \times \left(\frac{\max(N) + \min(N) - N_{patient}}{\max(N) + 1} \right) \tag{1}$$

$$\delta_{balanced} = \delta_{orig} \times \gamma_{patient} \tag{2}$$

By having this semi-supervised training strategy (supervised on the data feeding and unsupervised on the parameters updating), SciCNN can resemble different EEG patterns on a patient-independent feature space based on the knowledge of seizure/normal labeling (Fig. 8). Fig. 9 shows the t-SNE visualization of the 16-D feature space. The two subfigures show the locations of the NPCs trained with different groups of patients. Red dots represent the seizure NPCs. Blue dots represent the non-seizure NPCs. Gray dots are the EEG patterns that did not manifest on the specific patient. Nevertheless, they are still necessary for other patients who have different seizure patterns.

During the actual deployment of the trained SciCNN to the patient who is totally new to the classifier, a short (2-min) automatic on-chip calibration is conducted at the start to label the NPCs based on the normal EEG patterns of the target patients. Since

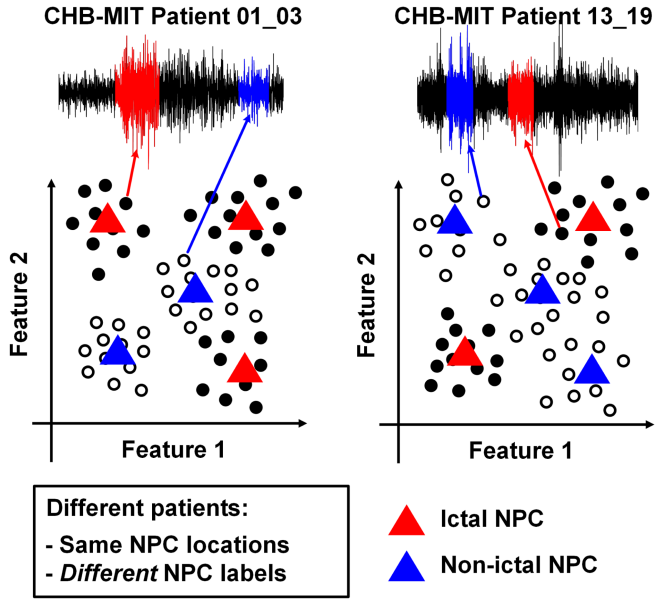

Fig. 8. Illustration of inference of Neural Pattern Clustering (NPC).

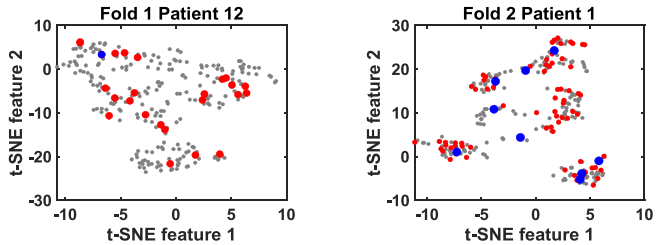

Fig. 9. Visualization of reduced dimensionality of neural pattern clusters with patient 12 and patient 1 in CHB-MIT database. .

the calibration is conducted when the patient is seizure-free, all the NPCs that have been identified as the closest NPCs will be labeled as non-seizure NPC. To reduce the false negative rate, these NPCs will only be labeled after having >2 (programmable) feature points that lie the closest to them.

C. Ring Global Buffer (RGB)

This work implements three Ring Global Buffers (RGBs), 2KB SRAM each, to store the digitalized data sent from the decimation unit. Two decimation blocks are implemented to cater to surface EEG and iEEG, respectively. With 20-tap FIR filters, one decimation unit decimates the input signal from 256

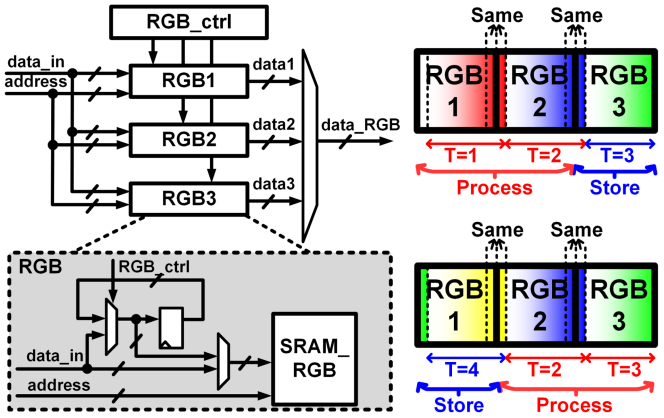

Fig. 10. Storing and processing sequence of ring global buffer (RGB).

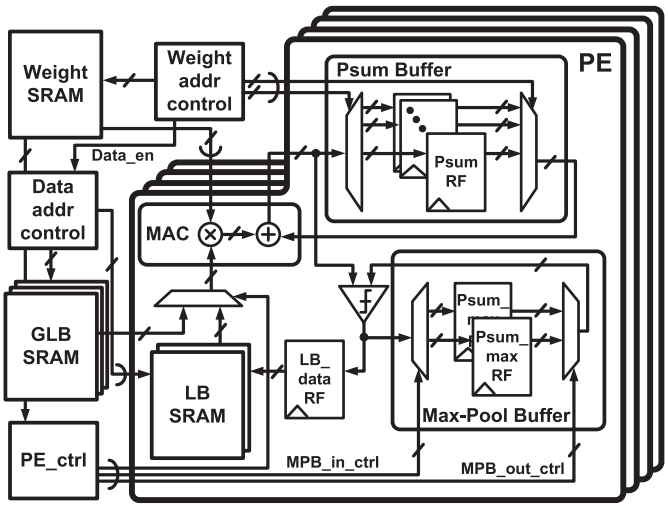

Fig. 11. Schematic of kernel-wise pipeline (KWP).

Hz to 64 Hz; another decimates from 1024 Hz to 128 Hz. The ADC sampling rate is adjustable based on the EEG electrode types. After decimation, each RGB is controlled separately to store 1 s/0.5 s data each for EEG/iEEG signals (Fig. 10). During the normalization state, mean and standard deviation of the 3 s/1.5 s EEG/iEEG data are computed.

$$\mu_{overall} = \frac{\sum_{i=1}^{N} \mu_i}{N} \tag{3}$$

$$\sigma_{overall} = \sqrt{\frac{\sum_{i=1}^{N} \sigma_i^2}{N} + \frac{1}{N-1} \sum_{i=1}^{N} (\mu_i - \mu_{overall})^2} \tag{4}$$

Equation (3) and (4) show the equations to compute the overall mean and overall standard deviation, respectively. N shows the number of rounds to compute the partial mean, μ_i, and the partial standard deviation, σ_i. By computing the partial mean and partial standard deviation, long period of data can be split into multiple 3s partial windows and processed sequentially. As a result, three GLBs that store 1s each can be reused for the computation of the mean and standard deviation of longer period of data. To further reduce the hardware complexity, 10b fixed-point adders are used for calculating the mean, where the division is done by a right shifter (shift by 3 bits to divide 24s data). To calculate the standard deviation, a 16b floating-point multiplier is used for higher precision. After the overall mean and the overall standard deviation are obtained, the decimated data will be normalized using these calculated mean and standard deviation, followed by being stored into RGBs.

During the actual process, each RGB takes turns to receive and store the normalized data. Two fully stored RGBs provide SciCNN with the normalized data as input; while the remaining one receives and stores the latest data. This ring structure realizes 2 s/1 s data processing with 1 s/0.5 s overlapping for CHB-MIT/EU database. To filter the normalized data for the depth dimension of SciCNN input, seven 20-tap FIR filters are designed. For CHB-MIT database, seven 4 Hz-bandwidth bands ranging between 0–28 Hz frequency are used; for EU database, six 4 Hz-bandwidth bands ranging between 0–24 Hz frequency and one highpass filter (>24 Hz) are used. Since causal filter requires previous data points for correctly filtering the current data point, the last few data points of the previous sample window are stored in two RGBs, one is the RGB that is storing the latest data and another one is the RGB next to it (Fig. 10). No data collision will happen as all the SciCNN computations finish in 0.59 s with the system clock of 4 MHz.

D. Kernel-Wise Pipeline (KWP)

Hardware-efficient data movement is essential for realizing a low-power CNN accelerator [27]. Among the power consumption of the data management within the chip, SRAM access dominates [28]. General-purpose CNN accelerators thrive to realize hardware-efficient data movement while processing the neural network computation. However, they normally cater to multiple types of popular neural networks (AlexNet, VGGNet, GoogLeNet, and ResNet) and utilize fast clock speed (>100 MHz) for high throughput (>1 GOPS). The resulting complicated logics control and high power consumption (>100 mW [29], [30]) are hence not optimal for the biomedical wearable devices. Moreover, they usually need to associate with off-chip memory for storing large amount of input data, which will become a hindrance for the users if the device footprint increases [31]. To customize to the patient-independent seizure tracking SoC implemented with SciCNN, Kernel-Wise Pipeline (KWP) is utilized to achieve low-power neural network processing on an Application Specific Integrated Circuit (Fig. 11).

$$O_{w,m} = B_m + \sum_{d=1}^{D} \sum_{l=1}^{L} W_{l,d,m} I_{w+l,d} \tag{5}$$

$$N_{comp} = \sum_{w=1}^{W} \sum_{m=1}^{M} O_{w,m} \tag{6}$$

Equation (5) shows the computation of $O_{w,m}$, which is the output feature map (ofmap) at location w on the input feature map (ifmap) computed with filter m of length L and depth D. I is the ifmap, W is the weight, B is the bias. (6) shows the number of computations N_{comp} to yield one row (one sensor) of ofmap with width W and stride of 1. It can be observed that

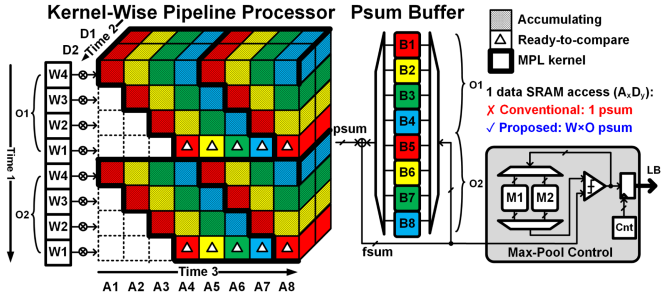

Fig. 12. Signal flow of kernel-wise pipeline (KWP), sizes of the kernels and the buffers have been reduced for the ease of visualization.

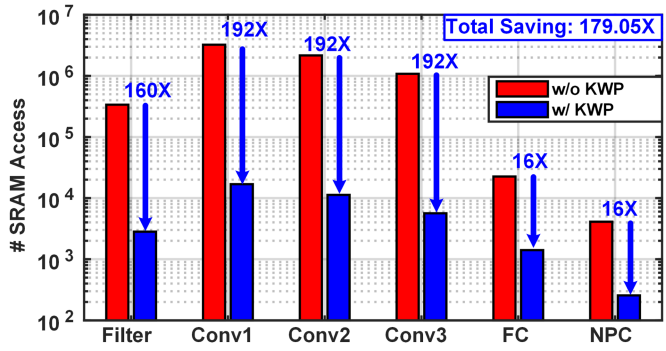

Fig. 13. Data SRAM saving by KWP.

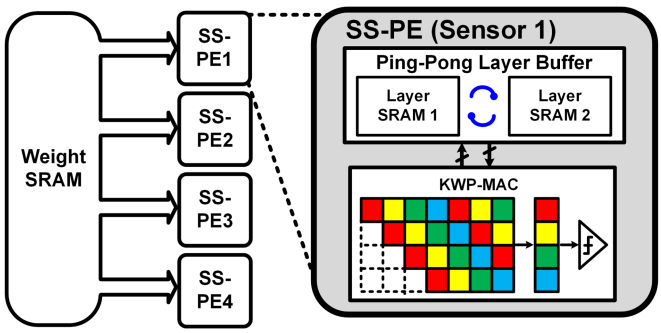

Fig. 14. Sensor-stationary process elements (SS-PE).

two adjacent ofmaps will reuse $L-1$ ifmaps, hence resulting in re-reading the same addresses in data SRAM for $M(L-1)$ times. To compute all the ofmaps in one zero-padded row, the repeating data SRAM access adds up to 99.4% of the total data SRAM access in our proposed structure. To alleviate this power-consuming SRAM access, Kernel-Wise Pipeline (KWP) is utilized to avoid accessing data SRAM for the overlapped ifmaps.

Fig. 12 illustrates the data movement of KWP. After one ifmap is read from data SRAM (A_xD_y), it will be reused $L \times O$ times to compute all the involving ofmaps in all branches. Corresponding weights (W_aO_b on Time1 direction) are read from weight SRAM sequentially to compute the partial sum (psum) of each ofmap, which is stored into Psum Buffer (PB) at different addresses. PB is built with 256-word 16b register files to accommodate the largest value of $L \times O$ in all the network layers without overflow.

To avoid storing all ofmap before starting to process Maxpool Layer (MPL), MPL is processed in a pipeline fashion and is facilitated by reading ifmap along D-dimension (Time2) first, followed by A-dimension (Time3). When an ofmap is fully computed (fsum), positive ofmap will be compared against the Ofmap Candidate (OC) stored in Maxpool Buffer (MPB), which is set to be equal to ofmap at the start of every Maxpool Kernel (MPK). MPB is built with 32-word 16b register files to accommodate the largest number of filters used in all SciCNN layers. OC will be subsequently updated with the larger value after each comparison of the fully-computed ofmap. On the other hand, negative ofmap is skipped and the comparison is data-gated. Max-pooled ofmap (mp-ofmap) is obtained every M comparisons, which will be stored into Ping-Pong Layer Buffer (PP-LB) as ifmap of the next layer.

To give an example, when A_4D_1 is read, the psum $A_4O_1W_4D_1$ is computed and stored in B_4. Then, without reading the next ifmap, $A_4O_1W_3D_1$ is computed and stored in B_3. This continues until $A_4O_2W_1D_1$ is computed and stored in B_5. Next, A_4D_2 is read, and the same process repeats. When $A_4O_1W_1D_2$ is computed (triangle tile), all the psum of the first ofmap is fully computed and accumulated (fsum). Since it is the first ofmap of the first MPL, M_1 updates into this value. When the last psum of the second ofmap is computed ($A_5O_1W_1D_2$), if it is a positive value, it will be compared against M_1. Then, M_1 updates into the larger value. This goes on until M_1 has compared against $A_7O_1W_1D_2$ and updated accordingly. The final M_1 value is sent to Ping-Pong Layer Buffer (PP-LB) for storing. Fig. 13 shows the reduced data SRAM access rates in each network layer, which sums up to 179.05× saving in total.

E. Ping-Pong Layer Buffer (PP-LB)

After performing the intra-layer acceleration with KWP, intermediate data on the inter-layer level (ifmap and ofmap) need to be stored efficiently to minimize the area consumption. In this work, we implement Ping-Pong Layer Buffer (PP-LB) to manage the intermediate data. PP-LBs consist of two SRAMs. At the first SciCNN layer, ifmap is read from GLB and ofmap is stored in one PP-LB. At the subsequent SciCNN layers, ifmap is read from the PP-LB that stores the ofmap of the previous layer; meanwhile, ofmap of the current layer is stored in another PP-LB. Hence, the PP-LBs are sized according to the ofmap sizes of the first two layers, which yield the largest numbers (12KB and 6KB) among all the layers. PP-LB successfully reduces the required SRAM size by 23.67%.

F. Sensor-Stationary Process Elements (SS-PEs)

Elevating from the localized level of the intra-PE processing, inter-PE data movement is also accelerated. Thanks to the 1D filter kernels used by all the SciCNN layers, all the rows in ifmap can be processed independently. Hence, we implement four Sensor-Stationary Process Elements (SS-PEs) to process the rows in parallel. After receiving the signals recorded from different sensors on the brain, four SS-PEs process 22 sensor channels by tiling them into four groups (Fig. 14). Thanks to the parallel processing of the SS-PEs, the total processing latency

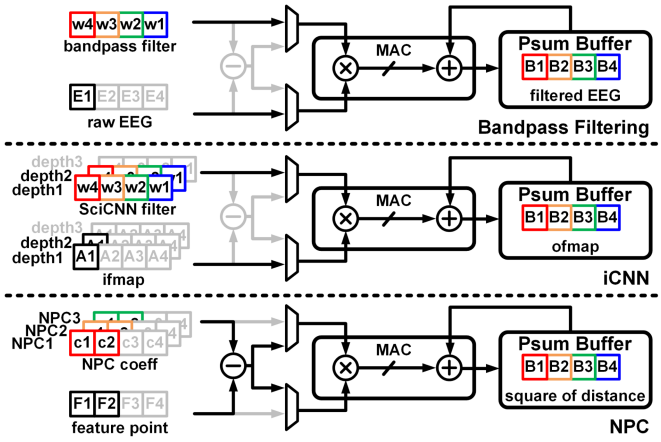

Fig. 15. Hardware reuse among bandpass filtering, iCNN, and NPC.

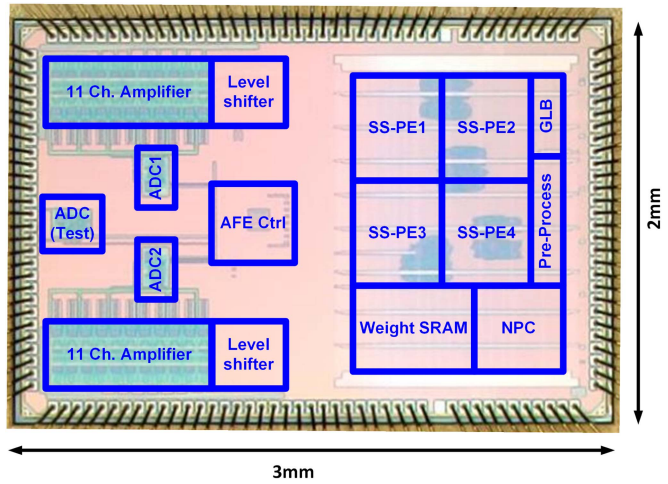

Fig. 16. Chip micrograph.

is reduced by ∼4×. Furthermore, as all SS-PEs process their ofmap in parallel, weight value is broadcasted to all the SS-PEs. This reduces weight SRAM access rate by ∼4×, hence reducing the power consumption proportionally.

G. Neural Pattern Clustering

Inter-patient variation is one of the biggest challenges while designing a patient-independent classifier. The characteristics of the seizure events recorded from different patients can be significantly different (Fig. 3). Furthermore, the characteristics of the seizure event of one patient can be well similar to the characteristics of the normal signal of another patient. Hence, Neural Pattern Clustering (NPC) is adopted to extract the neural pattern distribution.

After the coordinates of the NPCs have been fully trained with the existing databases, they are preloaded to the weight SRAM before deploying to the new patients. During the actual classification, ofmap of the last layer in inception-based CNN (iCNN) represents the coordinates of the feature point, which serves as the ifmap of NPC layer. Euclidean distances between the generated feature point and all NPC centroids are computed sequentially. Since we are interested in relative distances only, computational-expensive square root process is removed for power saving and area saving. Furthermore, the MAC unit and the adder are reused (Fig. 15), which is facilitated by Dimension-Wise Ofmap Pipeline (DW-OP). Here, each ifmap is the coordinate of the feature point in each dimension on the feature space. To save the data SRAM access rate, each ifmap value is read once-and-only-once. Then, psum is computed, which is equivalent to the partial sum of the distances to all the NPCs in one feature space dimension. PB is reused to store the psum of different NPCs. Finally, all the distances are fully computed when the last ifmap is read from data SRAM. Thanks to DW-OP, the data SRAM access rate is reduced by 16×. The stored distances are subsequently compared one-by-one to obtain the closest NPC. The final classification result is determined by the label of the closest NPC.

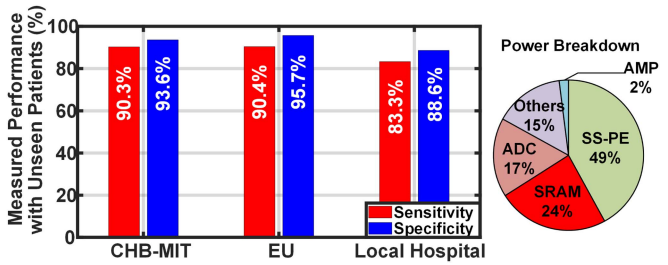

Fig. 17. Measured average event-based sensitivity, specificity, and power breakdown of SciCNN.

IV. Measurement Results

The proposed 0-shot-retraining patient-independent epilepsy detection SoC is fabricated in 40 nm 1P8M CMOS process. It integrates DDR-NQAFE and SNP in an area of 3 mm × 2 mm, where the active area per channel is 0.114 mm^2 (Fig. 16). Power breakdown is shown in Fig. 17. Table I shows the comparison between this work and the state-of-the-art epilepsy-tracking SoCs. This work, to the best of the authors' knowledge, is the first work that verifies the patient-independent seizure tracking SoC on both surface EEG (CHB-MIT database) and iEEG (EU database) and achieves event-based sensitivity of >90% and specificity of >93% on both databases. Although the performance drops compared to other arts due to the inter-patient variation, it advances in the following aspects:

1) Training the classifier without the need of pre-recording the EEG signals of the target patients, hence capable of being directly deployed to the new patients.
2) Performing unsupervised on-chip calibration that does not require the input from the patient.
3) General applicability has been verified by measuring the tape-out chip on 24 surface EEG subjects and 20 iEEG subjects, plus a local hospital patient.

A. Verification setup

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

TABLE I
COMPARISON WITH STATE-OF-THE-ART EPILEPSY MANAGEMENT SOCS

	This work			Hsieh JSSC'23 [11]	Zhang JSSC'22 [12]	Chua JSSC'22 [13]		Shin JSSC'22 [14]		Liu ISSCC'21 [6]	O'Leary ISSCC'20 [18]	Wang ISSCC'20 [5]
Process (nm)	40			40	40	28		65		65	65	180
Supply Voltage (V)	1.1 [AFE] / 0.9 [DBE]			0.49	0.7	0.5		1.2		0.75	1.2	1.5
Area/Ch (mm^2)	0.114			1.96	0.130	0.0125		0.0140		1.74	0.190	0.730
Channel	22			N/A	16	8		256		2	8	8
Algorithm	SciCNN			SVM	GTCA-SVM	LR		NeuralTree		CNN/MLP	EDM-DF	Coarse (Threshold) Fine (LS-SVM)
Patient-Independent	O			X	X	X		X		X	X	X
Validation (# Patients)	Local Hospital (1)	CHB-MIT (24)	EU (20)	CHB-MIT (24)	CHB-MIT (24)	CHB-MIT (24)	UoM (3)	CHB-MIT (24)	iEEG.org (6)	Bonn (N/A)	EU (N/A)	CHB-MIT (23)
Patient-Specific Training Sample %	0-Shot-Retraining			1/Few-Shot-Training								
	0%			35%	35%	30%		80%		N/A (>0%)	N/A (>0%)	N/A (>0%)
Sensitivity (%)	83.3	90.3	90.4	98.6	100	97.5	97.9	95.6	94.0	N/A	96.7	97.8
Specificity (%)	88.6	93.6	95.7	99.7	99.5	98.2	98.2	96.8	96.9		0.80 FP/hr	99.7
Accuracy (%)	86.0	91.9	93.0	99.2	99.8	97.9	98.1	96.2	95.5	99.8	N/A	98.75
Energy (µJ/Class)	28.33^{+}		21.63^{+}	0.096 (DBE only)	0.97	0.0015 (DBE only)		0.23		2.06 (DBE only)	0.036 (DBE only)	174^
Latency (s)	4.4	8.3	17.0	N/A	0.7	1.6		<1		N/A	N/A	<0.3

^estimated
surface EEG: Local Hospital, CHB-MIT, Bonn; iEEG: EU, UoM, iEEG.org, Bonn
$^{+}$Energy = Power / Sampling rate$^{\#}$
$^{\#}$Sampling rate: surface EEG = 64Hz, iEEG = 128Hz

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Accuracy = \frac{Sensitivity + Specificity}{2} \tag{9}$$

Event-based sensitivity and specificity are used in this work as the verification metrics, shown in (7) and (8), respectively. TP and TN are correctly-detected seizure event and correctly-classified normal samples, respectively; whereas FN and FP are missed seizure event and incorrectly-classified normal samples, respectively. Both event-based sensitivity and specificity are equally crucial in justifying the classification performance, as they show the ability in detecting the seizure events and avoiding false-alarms, respectively. Accuracy is shown with the average of sensitivity and specificity as an overall justification (9).

Terasic DE10-Nano FPGA equipped with Cyclone V SoC (Intel) is used in the setup to control the chip and provide 22 channels 10b data from the database to SNP at the sampling frequency of 64 Hz/channel and 128 Hz/channel for CHB-MIT database and EU database, respectively. Parameters (2017 bit) such as number of layers, number of inception modules in each layer, sizes of filter kernels, sizes of maxpool layers, and sizes of fully-connected layers are uploaded to the chip from FPGA at the start, followed by the trained weights of the filter kernels (73.13 KB).

B. CHB-MIT Database

The benefits of the 0-shot-retraining patient-independent epilepsy detection SoC is maximized when patients can directly use it without undergoing any surgery. Hence, open-access CHB-MIT database [32], [33] is used for verification, which contains 990 hours of *surface* EEG recordings and 180 seizure events from 24 pediatric patients (1.5–22 years old). All the seizure events and 18.38hr of non-ictal samples in total from the 24 patients are used to measure the event-based sensitivity and specificity with on-chip measurement. 8-fold cross verification is adopted in the verification. In each fold, SciCNN is trained only with 21 patients and is inferenced on the remaining three patients. In other words, the EEG samples of one patient are only assigned to either training set or testing set in each fold. This is to imitate the actual deployment that the trained device does not use any data from the target patient for the offline training. Table II shows the details of the patients grouping. All 22 channels that are commonly used in CHB-MIT database with the 10-20 system are adopted, namely T8P8, T7P7, T7FT9, P8O2, P7T7, P7O1, P4O2, P3O1, FZCZ, FT9FT10, FT10T8, FP2F8, FP2F4, FP1F7, FP1F3, F8T8, F7T7, F4C4, F3C3, CZPZ, C4P4, C3P3. By inferencing on the new patients with the 0-shot-retraining scheme, SciCNN successfully achieves *event-based* sensitivity and specificity of 90.3% and 93.6%, respectively, in average (Fig. 17). Detailed classification performance of each fold in the 8-fold cross validation is shown in Fig. 18.

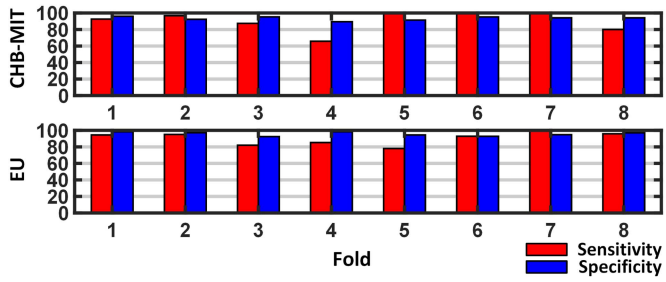

Fig. 18. Measured event-based sensitivity and specificity of SciCNN of all folds in 8-fold cross validation.

TABLE II
DETAILED INFORMATION OF 8-FOLD CROSS VALIDATION

Patient ID for Testing		
Fold	CHB-MIT (sample @64Hz)	EU (sample @128Hz)
1	12, 18, 20	10960, 11250
2	1, 11, 13	11460, 4420, 6350
3	16, 21, 24	10730, 10770, 130890
4	3, 14, 15	3840, 5900, 8180
5	7, 9, 19	9220, 10840, 11500
6	2, 8, 22	6200, 9580
7	4, 5, 17	5830, 132450
8	6, 10, 23	5480, 9160

# Tested Samples (hour)				
Fold	CHB-MIT		EU	
	Seizure	Normal	Seizure	Normal
1	0.21	1.98	0.73	1.93
2	0.13	2.48	1.81	3.16
3	0.10	1.49	0.97	5.45
4	0.35	2.40	2.47	2.67
5	0.11	2.39	1.15	3.74
6	0.17	1.50	0.26	1.35
7	0.16	1.84	0.10	1.59
8	0.14	4.30	1.99	2.27
Total	1.36	18.38	9.48	22.15

Latency (s)		
Fold	CHB-MIT	EU
1	11.53	22.42
2	3.19	24.79
3	4.68	33.12
4	13.98	23.01
5	3.22	7.70
6	5.54	8.90
7	20.87	5.39
8	3.50	10.27
Average	8.32	16.95

C. EU Database

To further verify the general applicability of SciCNN, SciCNN is verified with iEEG signals in European Epilepsia database (EU database) [34], [35], which contains >40000 hours iEEG recordings and 2400 seizure events. In this work, all 20 EU patients recorded with 1024 Hz sampling frequency are included in the verification. All the seizure events and 22.15hr of non-ictal signals are used in the verification during the on-chip measurement. 8-fold cross validation is also applied to split the 20 patients into training group and testing group (Table II). By training with iEEG signals with the 0-shot-retraining scheme, SciCNN successfully achieves 90.4% and 95.7% for event-based sensitivity and specificity, respectively (Fig. 17). Compared to CHB-MIT database, EU database yields higher specificity. It could be due to the lower recorded noise baseline, which is resulted by direct contact of the intracranial electrodes to the brain tissue. Besides, iEEG electrodes are less prone to motion artifact as well.

D. Local Hospital Patient

The general applicability to the new patients is further verified with a patient recruited from the local hospital. After training SciCNN with all the 24 CHB-MIT patients, a 1-hour recording from the local hospital's patient is inferenced. This recording includes 6 seizure events, each of which lasts for 7–11 s. As the results, SciCNN achieves 83.3% event-based sensitivity and 88.6% specificity, respectively. This result verifies that the performance of SciCNN is decent to apply to new patients while being trained with the pre-existing database. The reason that the local hospital's results are slightly inferior than the database results is because the EEG patterns from the sole recruited patient might be rather different from the NPC trained with the databases. This could be improved by either increasing the number of NPC or increasing the number of the training dataset.

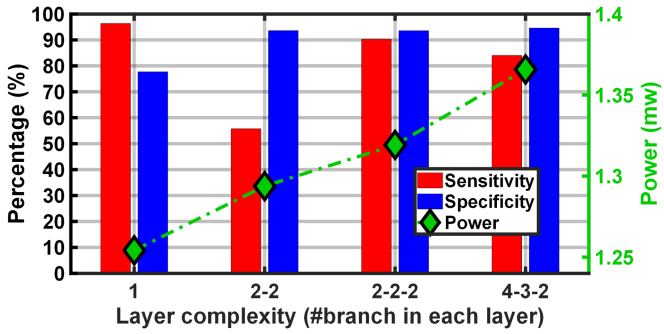

Fig. 19. Measured power performance with different layer complexity, each number in x-axis tick labels represents the number of branches in one layer.

E. Classifier Complexity

Multiple combinations of different numbers of layers and branches have been tested on the fabricated chip by changing the parameters uploaded to the chip at the start of the usage. Fig. 19 shows the power consumption and the classification performance with different layer complexity. As the complexity increases, the number of computations and power consumption increase proportionally. The corresponding classification performance in terms of event-based sensitivity and specificity plateaus after 3-layer 2-branch inception modules are used in iCNN. As the result, the combination of 3-layer iCNN with 2 branches in each layer is chosen (73.13KB), which yields >90% event-based sensitivity and >90% specificity simultaneously after inferencing with CHB-MIT database.

F. Discussion

Patient-specific classifiers have been extensively researched and have achieved excellent classification performance with optimized hardware efficiency. However, there is room for improvement in the practicality in terms of generalization. Patient-independent classifier can fill in this gap by skipping the part of data collection from the target patients. Researchers have proposed different algorithms, mainly deep learning classifiers, to achieve patient-independent classification. Zhu explored 0-shot and n-shot learning on the patient-independent feature space using a generative adversarial network-based classifier [36]. However, it is only verified with iEEG database, and it is less suitable for the actual hardware realization due to the complex structure of the deep learning classifiers. Li presented the parallel processing of memrister crossbar array that achieved both seizure detection and prediction with low latency and low power consumption [37]. However, it also lacks broader pools of databases for the verification, such as surface EEG databases that significantly increase the generalization of the application.

In this work, SciCNN is verified with both the surface EEG and iEEG databases with good classification performance (>90% event-based sensitivity and >95% specificity). Although they are relatively inferior to other patient-specific state-of-the-arts due to the inter-patient variation, it is important to know that

patients do not need to be hospitalized at all for any inpatient data collection. This highly increases the practicality of the epilepsy-tracking SoC, especially for the patients who cannot afford the high cost of the incurred treatment. Therefore, this work paves the way in delivering the automated machine-aided seizure-detection SoCs to more epileptic patients, while they can directly use the device with the least help from the professional. As the future direction, this work can serve as the fundamental classifier that performs preliminary on-chip classifications. The yielded classification results and the corresponding confidence level of the detection can be utilized in the follow-up online learning for updating the weights of the classifiers accordingly. In this way, unsupervised online learning can be achieved to further increase the vector-based sensitivity.

V. Conclusion

We present a 0-shot-retraining patient-independent epilepsy-tracking SoC that is verified with both surface EEG database and iEEG database, to the best of authors' knowledge, for the first time in the literature. In contrast to the conventional patient-specific seizure detection SoCs, the proposed SoC does not need to be trained with the EEG signals of the target patients, nor it requires to perform supervised online training after the actual deployment to the new patients. With the 0-shot-retraining patient-independent neural network structure and the hardware-efficient techniques, SciCNN achieves 90.3%/90.4%/83.3% event-based sensitivity and 93.6%/95.7%/88.6% specificity on unseen patients from CHB-MIT database/EU database/local hospital recruited patient, respectively.

References

[1] "Epilepsy," *World Health Organization*. 2019. Accessed: Jun. 20, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[2] L. Zhang, M. Zhang, C.-W. Tsai, and J. Yoo, "Review of AI-on-the-edge EEG-based patient-specific epilepsy tracking SoCs," in *Proc. IEEE 20th Int. New Circuits Syst. Conf.*, 2022, pp. 384–388.

[3] J. Li et al., "Body-coupled power transmission and energy harvesting," *Nature Electron.*, vol. 4, pp. 530–538, Jun. 2021, doi: 10.1038/s41928-021-00592-y.

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

[5] Y. Wang, Q. Sun, H. Luo, X. Chen, X. Wang, and H. Zhang, "A closed-loop neuromodulation chipset with 2-level classification achieving 1.5Vpp CM interference tolerance, 35dB stimulation artifact rejection in 0.5ms and 97.8% sensitivity seizure detection," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 404–405.

[6] J. Liu et al., "4.5 BioAIP: A reconfigurable biomedical AI processor with adaptive learning for versatile intelligent health monitoring," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2021, pp. 62–64.

[7] S. Y. Lee, Y.-W. Hung, Y.-T. Chang, C.-C. Lin, and G.-S. Shieh, "RISC-V CNN coprocessor for real-time epilepsy detection in wearable application," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 4, pp. 679–691, Aug. 2021.

[8] S. Zhao et al., "A 0.99-to-4.38 uJ/class event-driven hybrid neural network processor for full-spectrum neural signal analyses," *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 3, pp. 598–609, Jun. 2023.

[9] G. O'Leary, D. M. Groppe, T. A. Valiante, N. Verma, and R. Genov, "NURIP: Neural interface processor for brain-state classification and programmable-waveform neurostimulation," *IEEE J. Solid-State Circuits*, vol. 53, no. 11, pp. 3150–3162, Nov. 2018.

[10] C.-W. Tsai, M. Zhang, L. Zhang, and J. Yoo, "A closed-loop brain-machine interface with one-shot learning and online tuning for patient-specific neurological disorder treatment," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Circuits Syst.*, 2022, pp. 186–189.

[11] Y.-Y. Hsieh, Y.-C. Lin, and C.-H. Yang, "A 96.2nJ/class neural signal processor with adaptable intelligence for seizure prediction," *IEEE J. Solid-State Circuits*, vol. 58, no. 1, pp. 167–176, Jan. 2023.

[12] 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, Apr. 2022.

[13] A. Chua, M. I. Jordan, and R. Muller, "SOUL: An energy-efficient unsupervised online learning seizure detection classifier," *IEEE J. Solid-State Circuits*, vol. 57, no. 8, pp. 2532–2544, Aug. 2022.

[14] U. Shin et al., "NeuralTree: A 256-channel 0.227-μJ/class versatile neural activity classification and closed-loop neuromodulation SoC," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3243–3257, Nov. 2022.

[15] C.-W. Tsai et al., "SciCNN: A 0-shot-retraining patient-independent epilepsy-tracking SoC," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2023, pp. 488–489.

[16] R. Jiang, H. Wu, K. A. Ng, C.-W. Tsai, and J. Yoo, "A 13-bit 70MS/s SAR-assisted 2-bit/cycle cyclic ADC with offset cancellation and slack-borrowing logic," in *Proc. IEEE Eur. Solid-State Circuits Conf.*, 2023, pp. 281–284.

[17] M. A. Bin Altaf and J. Yoo, "A 1.83 μJ/classification, 8-channel, patient-specific epileptic seizure classification SoC using a non-linear support vector machine," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 1, pp. 49–60, Feb. 2016.

[18] G. O'Leary et al., "26.2 a neuromorphic multiplier-less bit-serial weight-memory-optimized 1024-tree brain-state classifier and neuromodulation SoC with an 8-channel noise-shaping SAR ADC Array," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2020, pp. 402–404.

[19] M. Shoaran, B. A. Haghi, M. Taghavi, M. Farivar, and A. Emami-Neyestanak, "Energy-efficient classification for resource constrained biomedical applications," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 4, pp. 693–707, Dec. 2018.

[20] A. Uran et al., "A 16-channel neural recording system-on-chip with CHT feature extraction processor in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 57, no. 9, pp. 2752–2763, Sep. 2022.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[23] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* ., 2016, pp. 770–778.

[26] S. Mark, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[27] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[28] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2014, pp. 10–14.

[29] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.

[30] H. Mo et al., "A 12.1 TOPS/W quantized network acceleration processor with effective-weight-based convolution and error-compensation-based prediction," *IEEE J. Solid-State Circuits*, vol. 57, no. 5, pp. 1542–1557, May 2022.

[31] M. Kim and J.-S. Seo, "An energy-efficient deep convolutional neural network accelerator featuring conditional computing and low external memory access," *IEEE J. Solid-State Circuits*, vol. 56, no. 3, pp. 803–813, Mar. 2021.

[32] A. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000.

[33] A. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. Thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, Sep. 2009.
[34] M. Ihle et al., "EPILEPSIAE - A European epilepsy database," *Comput. Methods Prog. Biomed.*, vol. 106, no. 3, pp. 127–138, Jun. 2012.
[35] J. Klatt et al., "The EPILEPSIAE database: An extensive electroencephalography database of epilepsy patients," *Epilepsia*, vol. 53, no. 9, pp. 1669–1676, 2012.
[36] B. Zhu and M. Shoaran, "Unsupervised domain adaptation for cross-subject few-shot neurological symptom detection," in *Proc. IEEE/EMBS 10th Int. Conf. Neural Eng.*, 2021, pp. 181–184.
[37] C. Li, C. Lammie, X. Dong, A. Amirsoleimani, M. R. Azghadi, and R. Genov, "Seizure detection and prediction by parallel memristive convolutional neural networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 4, pp. 609–625, Aug. 2022.

Chne-Wuen Tsai (Member, IEEE) received the B.E. degree in biomedical engineering in 2018 from the National University of Singapore, Singapore, where he is currently working toward the Ph.D. degree in electrical and computer engineering. His research interests include energy-efficient machine learning algorithm for biomedical wearable devices and resource-constrained system-on-chip design.

Rucheng Jiang (Member, IEEE) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2014, and the M.S. degree in electronic science and technology from Zhejiang University, Hangzhou, China, in 2017. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. From 2017 to 2020, he worked in industry in the design of high-performance Operational Amplifier (OPAMP) and voltage regulators. His research interests include energy-efficient high-speed and high-performance analog-to-digital converters.

Lian Zhang (Student Member, IEEE) received the B.E. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2016, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2021. She was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore afterwards. Since 2022, she has been with Apple, Cupertino, CA, USA. Her research interests include low-power and low-noise analog instrumentation, area- and energy-efficient data conversion, and mixed-signal system-on-chip (SoC) design for healthcare applications.

Miaolin Zhang (Student Member, IEEE) received the B.Eng. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, the distinguished graduate Diploma degree from the University of Strathclyde, Glasgow, U.K., in 2017, and the Ph.D. degree from the National University of Singapore, Singapore, in 2021. He is currently working with Huawei Technologies, Chengdu. His research interests include machine learning algorithm developing for resource-constrained hardware, energy-efficient SoC design, and ASICs for high-performance Ethernet switch.

Jerald Yoo (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2002, 2007, and 2010, respectively. From 2010 to 2016, he was with the Department of Electrical Engineering and Computer Science, Masdar Institute, Abu Dhabi, UAE, where he was an Associate Professor. From 2010 to 2011, he was a Visiting Scholar with the Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA, USA. Since 2017, he has been with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, where he is currently an Associate Professor. He has pioneered research on low-energy body area networks for communication/powering and wearable body sensor networks using the planar-fashionable circuit board for a continuous health monitoring system. He has authored book chapters in *Biomedical CMOS ICs* (Springer, 2010), *Enabling the Internet of Things—From Circuits to Networks* (Springer, 2017), *The IoT Physical Layer* (Springer, 2019), and *Handbook of Biochips (Biphasic Current Stimulator for Retinal Prosthesis)* (Springer, 2021). His research interests include low-energy circuit technology for wearable bio-signal sensors, flexible circuit board platform, BAN communication and powering, application-specific integrated circuits for piezoelectric micromachined ultrasonic transducers, and system-on-chip design to system realization for wearable healthcare applications. Dr. Yoo is/was a Technical Program Committee Member of the IEEE International Solid-State Circuits Conference, the Co-Chair of ISSCC Student Research Preview, and Emerging Technologies and Applications Subcommittee Chair of IEEE Asian Solid-State Circuits Conference and IEEE Custom Integrated Circuits Conference. He is also an Analog Signal Processing Technical Committee Member of the IEEE Circuits and Systems Society. He was the recipient or co-recipient of several awards, including the IEEE International Solid-State Circuits Conference (ISSCC) 2020 and 2022 Demo Award (Certificate of Recognition), IEEE International Symposium on Circuits and Systems (ISCAS) 2015 Best Paper Award (BioCAS Track), ISCAS 2015 Runner-Up Best Student Paper Award, Masdar Institute Best Research Award in 2015, and IEEE Asian Solid-State Circuits Conference (A-SSCC) Outstanding Design Award in 2005. He was the Founding Vice-Chair of IEEE SSCS United Arab Emirates (UAE) Chapter and Chair of the IEEE SSCS Singapore Chapter. He was a Distinguished Lecturer for the IEEE Circuits and Systems Society from 2019 to 2021 and IEEE Solid-State Circuits Society from 2017 to 2018.