2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9872006

Tiny CNN for Seizure Prediction in Wearable Biomedical Devices

Yang Zhang[1], Yvon Savaria[1], Shiqi Zhao[2], Gonçalo Mordido[1,3],
Mohamad Sawan[1,2], François Leduc-Primeau[1]

Abstract— **Epilepsy is a life-threatening disease affecting millions of people all over the world. Artificial intelligence epileptic predictors offer excellent potential to improve epilepsy therapy. Particularly, deep learning models such as convolutional neural networks (CNN) can be used to accurately detect ictogenesis through deep structured learning representations. In this work, a tiny one-dimensional stacked convolutional neural network (1DSCNN) is proposed based on short-time Fourier transform (STFT) to predict epileptic seizure. The results demonstrate that the proposed method obtains better performance compared to recent state-of-the-art methods, achieving an average sensitivity of 94.44%, average false prediction rate (FPR) of 0.011/h and average area under the curve (AUC) of 0.979 on the test set of the American Epilepsy Society Seizure Prediction Challenge dataset, while featuring a model size of only 21.32kB. Furthermore, after adapting the model to 4-bit quantization, its size is significantly decreased by 7.08x with only 0.51% AUC score precision loss, which shows excellent potential for hardware-friendly wearable implementation.**

I. Introduction

Nearly 60 million people in the world suffer from epilepsy, a common and serious brain disease which can affect people of all ages [1]. Epilepsy is characterized by unprovoked seizures, and can cause other health problems [2], which may be life-threatening. Long-time medication is a common method to control epilepsy, which can cause some undesirable side effects such as medication resistance [3]. It is especially important for epileptic patients to know when a seizure will happen to allow taking suitable mitigation measures in advance. To this end, seizure prediction can help them improve their well-being.

In clinical practice, brain electrical activities can be measured by multiple channels through a collection of electrodes installed on the scalp or exposed surface of the brain to collect scalp Electroencephalogram (sEEG) signals and intracranial EEG signals, respectively [4]. Thus long-term EEG monitoring to process neural signals is critical to patients due to the chronic characteristic of epilepsy. Hence, low-power acquisition and processing of EEG signals should be considered for wearable biomedical devices as well as implanted devices.

Seizure prediction is usually viewed as a binary classification problem between the preictal and non-preictal classes [3]. As shown in Fig. 1, the preictal state is the period

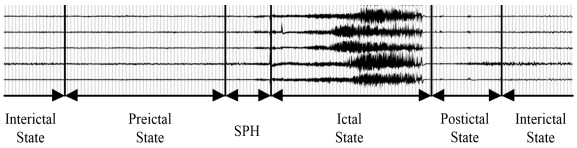

Figure 1. An example of multi-channel EEG recording for canine. Interictal state means seizure-free period, whereas ictal relates to the seizure onset period. The preictal state is the pre-seizure period, while the postictal state is the post-seizure period. The seizure prediction horizon (SPH) represents a period between preictal state and ictal state, where it is ideal to execute medical intervention or to apply risk mitigation measures.

before a seizure onset, while a non-preictal state can be one of three states: interictal (seizure-free), ictal (during a seizure) and postictal (after seizure). The main challenge of the seizure prediction problems is classifying signals into preictal and interictal states [3] with high sensitivity and low false alarm rate, which validates prediction while minimizing disturbance to the normal patients activities [5]. To this end, the recent development of promising deep learning techniques enabled significant performance improvements of seizure prediction methods. Truong et al. [6] developed a generalized retrospective and patient-specific seizure prediction method using STFT and CNNs, with average sensitivity of 75% and average false prediction rate (FPR) of 0.21/h on the American Epilepsy Society seizure prediction challenge dataset. Zhao et al. [7] explored energy-efficient seizure prediction, through feeding a direct end-to-end time-domain signal to a CNN, with an average sensitivity of 93.48%, average FPR of 0.063/h, and average AUC of 0.977 on the same dataset. Liu et al. [8] proposed a multi-view CNN to predict seizures, by combining time domain and frequency domain features in a CNN, with an average AUC of 0.837 on the same dataset.

Although, deep learning methods often achieve state-of-the-art results compared to traditional machine learning approaches, power consumption is significantly increased due to the size of existing deep learning models. By analyzing the trade-off between power consumption and performance, we propose means to reduce energy consumption while minimizing performance losses. To achieve this, a hardware-friendly tiny CNN for epileptic seizure prediction is proposed. Our main contributions can be summarized as follows:

1) A one-dimensional stacked CNN (1DSCNN) is proposed to predict epilepsy seizure for wearable biomedical devices. The proposed method outperforms existing methods in spite of a very competitive small model size.

2) Various quantization schemes are applied to the proposed 1DSCNN model for evaluating the impact of different bit widths on model performance.

The remainder of this paper includes the description of the

[1] Department of Electrical Engineering, Polytechnique Montreal, Canada
[2] School of Engineering, Westlake University, Hangzhou, Zhejiang, China
[3] Mila - Quebec AI Institute, Montreal, Canada
Corresponding author's e-mail: yang.zhang@polymtl.ca

TABLE I. PER-SUBJECT CHARACTERISTICS OF THE DATASET: NUMBER OF CHANNELS, SIZE OF THE PREICTAL SEGMENTS, SIZE OF THE INTERICTAL SEGMENTS, AND INTERICTAL HOURS

Subject	Channels	Preictal segments	Interictal segments	Interictal hours
Dog 1	16	24	480	80.0
Dog 2	16	42	500	83.3
Dog 3	16	72	1440	240.0
Dog 4	16	97	804	134.0
Dog 5	15	30	450	75.0
Patient 1	15	18	50	8.3
Patient 2	24	18	42	7.0

adopted methodology in Section II, our results are reported and discussed in Section III, finally, our main findings and conclusions are reported in Section IV.

II. METHODOLOGY

A. Dataset

In this work, the challenging and widely used dataset provided during the American Epilepsy Society seizure prediction challenge [9] is adopted as the benchmark dataset to compare various prediction methods. The dataset consists of iEEG recordings from five dogs and two human patients with naturally occurring epilepsy. Details about the collected information using an ambulatory monitoring system are shown in Table I. More specifically, recordings from four of the dogs are obtained through 16 subdural electrodes and the remaining dog through 15 electrodes, sampled at 400 Hz. One patient is recorded through 15 subdural electrodes, while the other patient through 24 electrodes, sampled at 5000 Hz. These are long-duration recordings, spanning from multiple months and up to a year. Also, in this dataset, interictal data segments are required to be at least one week before or after any seizure, while preictal data segments cover one hour before a seizure with a five-minute seizure horizon. The annotated dataset is divided into two parts for each subject through five-fold stratified cross-validation [10]: 80% training set and 20% testing set. Then, 20% of the training set is used as validation set.

B. Preprocessing

The preprocessing stage consists of data segmentation, resampling, and a short-time Fourier transform (STFT). During data segmentation, each 10-minute preictal or interictal segment is sliced into 20-second clips without overlap to augment the dataset. Then, each 20-second clip is resampled at 400 Hz for convenient processing. Before feeding the data to CNN, the initial raw iEEG data is converted into a two-dimensional time-frequency representation. Based on the non-stationary nature of iEEG signals, which highly depend on time, STFT is employed to convert time-series EEG signals into time-varying frequency components [11].

The raw iEEG signal is converted into 24 frequency bands according to brain wave frequencies from 0.1 Hz to 190 Hz as shown in Fig. 2. More specifically, delta (0.1-2, 2-4 Hz), theta (4-6, 6-8 Hz), alpha (8-10, 10-12 Hz), beta (12-21, 21-30 Hz), low-gamma (30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100 Hz) and high-gamma (100-110, 110-120, 120-130, 130-140, 140-150, 150-160, 160-170, 170-180, 180-190 Hz) [12]. When applying STFT, there is a trade-off between time

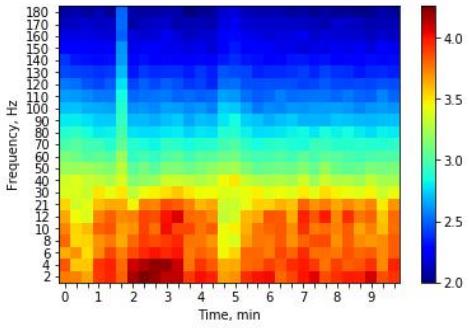

Figure 2. The mean value of spectrum amplitude in 24 frequency bands from 0.1 to 190 Hz of a 10-minute segment for a single channel.

resolution and frequency resolution. The 20-second window length is selected to guarantee a frequency resolution over 0.1Hz. A rectangular window shape is used because it reduces the main lobe width in the frequency domain, thus improving the frequency resolution [13]. Given an occurrence of abnormal brain discharge, the energy of the pre-seizure state is assumed to be concentrated in certain frequency bands. Hence, to reduce complexity and foster deployment on wearable biomedical devices, the mean value of the spectrum amplitude in each band is proposed as input features for the CNN. For each subject, the input consists of a 20-second clip, the CNN outputs one prediction for every time clip. The input size is $Number of\ channels \times 24$.

C. CNN Architecture and Training Settings

With the recent developments in bioinformatics, CNN is an attractive approach to analyze EEG signals [14] through the extraction of low-level features to be fed in subsequent layers to represent high-level features. In this work, a one-dimensional stacked CNN (1DSCNN) is proposed to predict epilepsy seizure. The overall CNN architecture is shown in Fig. 3. The stacked convolutional layer, initially proposed in VGGNet [15], presents two advantages: the depth of the neural network is improved and the amount of parameters is reduced under the condition of ensuring the same receptive field. Compared to two-dimensional CNN (2D CNN), one-dimensional CNN (1D CNN) can extract not only interior image pixels, but also more details about low-level features, such as edge shape, among multi-channels. As described in Fig. 3, firstly, 16-channel iEEG signals are passed through one 1DSCNN block to extract cross information between different channels at the same time. Then, two 1DSCNN blocks follow to improve the generalization of the model. The ReLU function is used for each layer. Finally, a Softmax layer follows to perform classification. The Adam optimizer is used for training, with a varying learning rate from 10^{-3} to 5^{-4}, β_1 and β_2 of 0.9 and 0.999 respectively. The learning rate is decreased if the validation error is not improved. Batch normalization and dropout are applied during training to prevent overfitting. The model in this work is implemented in Python 3.6 using Keras 2.3.1 with a Tensorflow 1.13.1 backend. The model is configured to run in parallel on two NVIDIA Tesla V100 graphics cards.

D. CNN Quantization

For meeting the time and energy constraints in wearable

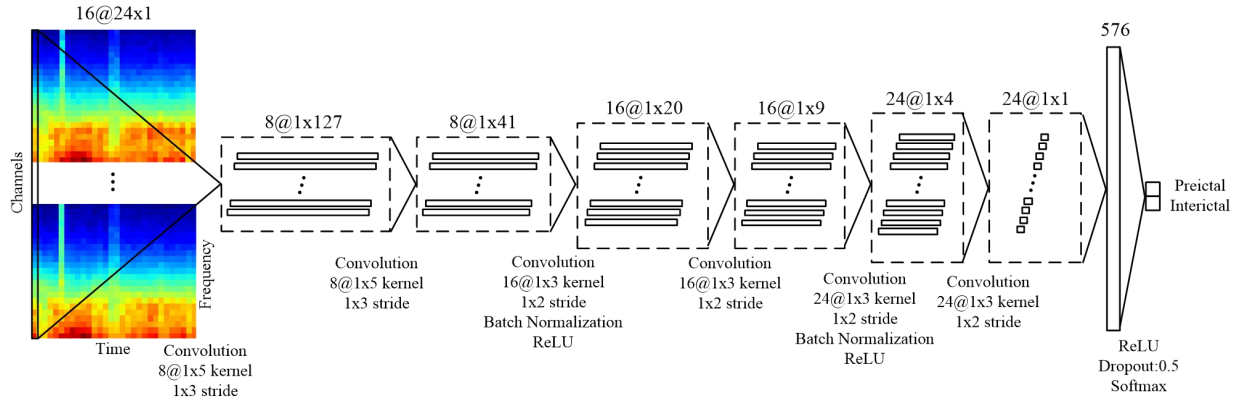

Figure 3. Architecture of the proposed convolutional neural network.

biomedical devices, we quantize the CNN weights and activations through re-training to reduce computation time, memory requirements, and power consumption [16]. We evaluate the impact of quantization on model performance using different bit widths. During the forward pass, weights and activations are quantized as fixed-point values with the same precision through uniform symmetric signed and uniform asymmetric signed quantization, respectively [17]. The scaling factor is a power of two, which allows the scaling to be computed using bit shifts instead of multipliers [18]. Moreover, the Tanh function is used instead of the ReLU function [7] due to the improvement of AUC scores. During the backward pass, the gradient is propagated by full-precision weights and straight-through estimators (STE) [19]. We note that no quantization is applied to the input and output layers.

III. Results and Discussion

A. Evaluation Metrics

A rigorous evaluation methodology is used to assess model performance on each subject of the dataset through five-fold cross-validation. Sensitivity, false prediction rate (FPR) and area under the ROC curve (AUC) are computed to evaluate our approach and compared with recent state-of-the-art works. Sensitivity is the percentage of correctly classified 20-second seizure clips among the total number of 20-second seizure clips. FPR is defined as the false positive rate per hour [6]. AUC is the area under the receiver operating characteristic curve (ROC), which illustrates the diagnostic ability of a given classifier.

B. Performance Analysis

Table II shows the evaluation results of the proposed 1DSCNN model, achieving an average sensitivity of 94.44%, an average FPR of 0.011/h, and an average AUC of 0.979 for all subjects on the dataset. Median and deviation values for AUC are presented in Fig. 4 through five-fold cross-validation, where the yellow line in the box and the edge of the box refer to median and quartile values of AUC, respectively. And the bar of box varies from minimum to maximum values of AUC. It is observed that Dog 1 is the subject for which seizures are hardest to predict, because even after tuning hyperparameters as best as we can, the AUC of Dog 1 remains the lowest among all subjects.

Table III compares our method with other state-of-the-art methods on the same dataset. Model size is reported for 32-bit full precision parameters. The comparison results demonstrate the proposed 1DSCNN model achieves the best sensitivity, FPR, and AUC at the lowest model size. It is worth noting that, although the average AUC score of our method is 0.002 higher than Zhao et al. [7], our model has less than half the size of their model, showcasing the potential of our model for wearable biomedical devices.

Fig. 5 demonstrates AUC scores and model size after different quantization levels. Compared to the full precision (FP) baseline, 8-bit quantization reduces model size by 3.79 times, with only 0.31% AUC score loss. Moreover, 4-bit quantization reduces model size by 7.08 times with only a

TABLE II. Per-subject evaluation results: sensitivity, false prediction rate (FPR) and area under the ROC curve (AUC)

Subject	Sensitivity(%)	FPR(/h)	AUC
Dog 1	91.11	0.013	0.926
Dog 2	97.70	0.001	0.998
Dog 3	95.42	0.003	0.978
Dog 4	92.27	0.003	0.974
Dog 5	96.78	0.001	0.999
Patient 1	97.22	0.008	0.998
Patient 2	90.56	0.045	0.979
Average	94.44	0.011	0.979

TABLE III. Comparsion with other state-of-the-art methods

Method	Sensitivity(%)	FPR(/h)	AUC	Model Size
Truong et al.[6]	75.00	0.210	-	0.76MB
Zhao et al.[7]	93.48	0.063	0.977	45.22kB
Brinkmann et al. [12]	-	-	0.860	-
Iryna et al. [3]	-	-	0.810	0.56MB
Liu et al. [8]	-	-	0.837	1.79MB
This work	94.44	0.011	0.979	21.32kB

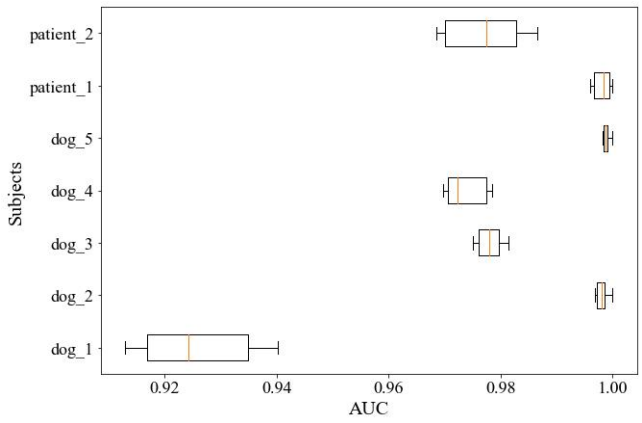

Figure 4. Per-subject AUC box plot: median and deviation of AUC scores through five-fold cross-validation.

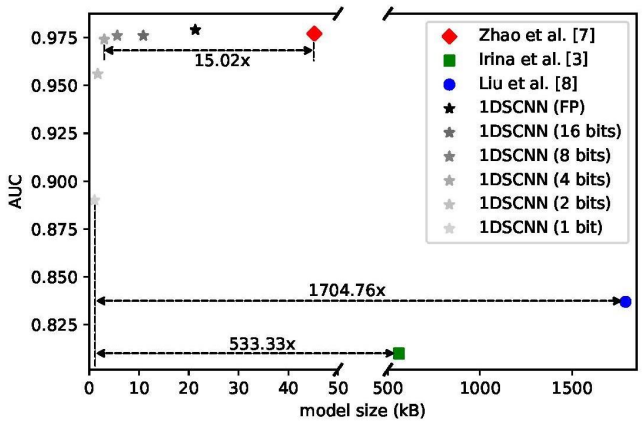

Figure 5. Comparison of the proposed methods with other state-of-the-art methods: AUC versus model size.

0.51% reduction in AUC score. Finally, 2-bit and 1-bit quantization reduce the model size by 12.54 and 20.30 times with 2.35% and 9.09% AUC score loss, respectively. Hence, 4-bit quantization shows great promise for wearable biomedical devices, significantly reducing the model size at a tolerable precision loss. Fig. 5 compares results obtained with 1DSCNN implemented at various precisions ranging from floating point to 1 bit with other state-of-the-art methods. Compared to Zhao et al. [7], a 4-bit quantized 1DSCNN reduces the model size 15.02 times with only 0.31% AUC score loss. More notably, compared to Liu et al. [8] and Irina et al. [3], 1-bit quantization significantly reduces by 1704.76 and 533.33 times the model size while offering 6.33% and 9.88% AUC score improvement, respectively. The reason why the model size of Liu et al. [8] is so large is that their model exploits two-dimensional inputs and employs more layers. While Irina et al. [3] adopts large receptive field filters and input size. These comparisons show the proposed methods outperform existing state-of-the-art methods.

IV. Conclusion

This paper proposes a one-dimensional stacked CNN (1DSCNN) for epilepsy seizure prediction with a model size suitable for wearable biomedical devices. Compared to recent state-of-the-art methods, the proposed 1DSCNN achieves the best performance with the lowest model size on the American Epilepsy Society Seizure Prediction Challenge dataset. When combined with quantization, our method is hardware-friendly, easing its deployment in wearable biomedical devices. Further work will consider advanced binary quantization methods of CNN to further improve the AUC of tiny models which are suitable for biomedical implanted devices.

Acknowledgments

The authors acknowledge financial support from IVADO (grant PRF-2019-4784991664), and the China Scholarship Council.

References

[1] Bandarabadi M. “Low-complexity measures for epileptic seizure prediction and early detection based on classification,” [Doctoral dissertation], 2015.

[2] Black M, Graham D I. “Sudden death in epilepsy,” Current Diagnostic Pathology, vol. 8, no. 6, pp. 365-372, 2002.

[3] Korshunova, Iryna, et al. “Towards improved design and evaluation of epileptic seizure predictors.” IEEE Transactions on Biomedical Engineering, vol. 65, no. 3, pp. 502-510, 2017.

[4] Yuan Y, Xun G, Jia K, et al. “A Multi-View Deep Learning Framework for EEG Seizure Detection,” IEEE journal of biomedical and health informatics, vol. 23, no. 1, pp. 83-94, 2019.

[5] Aarabi A, He B. “A rule-based seizure prediction method for focal neocortical epilepsy,” Clin. Neurophysiol, vol. 123, no. 6, pp. 1111-1122, 2012.

[6] Truong N D, Nguyen A D, Kuhlmann L, et al. “Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram,” Neural Networks, vol. 105, pp. 104-111, 2018.

[7] Zhao, Shiqi, Jie Yang, and Mohamad Sawan. “Energy-efficient neural network for epileptic seizure prediction.” IEEE Transactions on Biomedical Engineering, vol. 69, no. 1, pp. 401-411, 2021.

[8] Liu C-L, et al. “Epileptic Seizure Prediction with Multi-View Convolutional Neural Networks.” IEEE Access, vol. 7, pp. 170352-170361, 2019

[9] American Epilepsy Society Seizure Prediction Challenge, 2015. [Online]. Available: www.kaggle.com/c/seizure-prediction/data.

[10] Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

[11] Czarnek, Nicholas, et al. “The impact of time on seizure prediction performance in the FSPEEG database.” Epilepsy & Behavior, vol 48, pp. 79-82, 2015.

[12] Brinkmann, BH., et al. “Crowdsourcing reproducible seizure forecasting in human and canine epilepsy.” Brain, vol 139, pp. 1713-1722, 2016.

[13] Oppenheim, Alan V., John R. Buck, and Ronald W. Schafer. Discrete-time signal processing. Vol. 2. Upper Saddle River, NJ: Prentice Hall, 2001.

[14] Schirrmeister RT, et al. “Deep learning with convolutional neural networks for EEG decoding and visualization.” Human brain mapping, vol. 38, no. 11, pp. 5391-5420, 2017.

[15] Simonyan K, Zisserman A, “Very deep convolutional networks for large-scale image recognition”, International Conference on Learning Representations., pp. 1-14, 2015.

[16] X. Lin, C. Zhao, and W. Pan, “Towards accurate binary convolutional neural network,” Conference on Neural Information Processing System, pp. 345–353, 2017.

[17] Coelho, Claudionor N., et al. “Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors.” Nature Machine Intelligence, vol. 3, pp. 675-686, 2021.

[18] Moons, Bert, et al. “Minimum energy quantized neural networks.” 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 2017.

[19] Hubara, Itay, et al. “Quantized neural networks: Training neural networks with low precision weights and activations.” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 6869-6898, 2017.