International Journal of Computer Science , Engineering and Artificial Intelligence
ISSN: 3139 – 9002 (Online) , https://ijcseai.com/
Regular Issue, Vol. 3, Issue. 1 (January – March) , 2026 , Pages: 8 - 12
Received: 14 June 2025 ; *Accepted:* 25 December 2025 ; *Published:* 17 January 2026
https://doi.org/10.63328/IJCSEAI-V3RI1P2
https://www.youtube.com/watch?v=b7TXOaxLJqQ
https://archive.org/details/jsp-ijcseai-v3ri1p2
Original Paper : Qualitative Research – Analytical Research Paper

Real-Time Paediatric Seizure Monitoring on Edge Devices Using Lightweight Temporal Convolutional Networks

M. Dharani [1] , Nasreen Sultana Quadri [2] , T. Venkata Krishnamoorthy [3]

[1] Department of ECE, Mohan Babu University, Tirupati, India
[2] Dept of computer science and information, College of science, Zulfi Majmaah University, Saudi Arabia.
n.quadri@mu.edu.sa
[3] Department of ECE, Sasi Institute of Technology and Engineering, Tadepallegudem, Andhra Pradesh, India;
krishna454@gmail.com

* *Corresponding Author: M. Dharani ; dharani405@gmail.com*

Abstract: Epilepsy among children is one of the most common neurological conditions with a need for ongoing monitoring for prompt clinical intervention. Using conventional cloud based systems can lead to time delays and be problematic for privacy reasons and therefore are inappropriate for providing real-time pediatric care. A real time monitoring framework deployed in resource constrained edge devices is proposed in this paper. The architecture comprises of wearable EEG acquisition, bandpass, PCA-based feature extraction and then optimises the lightweight Temporal Convolutional Network (TCN) for deployment on TensorFlow Lite using model pruning and quantization. A classification accuracy of 97.6%, 70.8% lower than the cloud rely CNN approach, AUC of 0.987, and inference latency of 38ms are achieved with only 95mW being consumed during the experiment with the pediatric EEG benchmark. This finding shows that such on-device and privacy-preserving low latency seizure monitoring is possible for young patients.
Keywords: Pediatric Epilepsy, EEG, Edge Computing, TinyML, TCN, real-time seizure detection; wearable systems

1. Introduction

EEG is the fourth major disorder of the brain after stroke, dementia, Parkinson's disease and traumatic injuries and affects children particularly. About 1 child out of every 100 years has epilepsy and uncontrolled fits can lead to cognitive problems, injuries and sudden unexpected death in epilepsy (SUDEP) [1]. While in-hospital continuous electroencephalogram (EEG) monitoring is currently the standard for detecting seizures, the change of monitoring from inpatient to ambulatory is accompanied by many engineering challenges. Conventional hospital based EEG devices are large and value rigged and require child discipline to stay still. In recent years, there have been proposals to increase the coverage of the monitoring using cloud-computing-based solutions that create a 100–300 ms latency, huge bandwidth requirement and privacy issues for the sensitive neurological data [3]. These restrictions are especially inconvenient in the field of paediatrics, where it is essential for an alarm to be issued quickly and information is confidential. Both edge computing and TinyML are becoming new paradigms that move the inference process from its central location to the device itself. Quantized deep-learning (DEDL) models can be executed swiftly, with only sub-100 mW power consumption, in tens of milliseconds on the microcontroller or embedded SoC with neural processing units (NPUs), allowing for consistent and privacy-safe monitoring [3]. It is not trivial, however, to design a sufficiently accurate, yet compactly sized model for pediatric EEG, as EEG morphology varies widely across children, tolerable variations to the placement of the electrodes is limited, and the memory capacity of edge hardware remains limited.

The main contributions of this paper are: (1) a full pipeline of seizure detection validated on pediatric EEG recordings that can be deployed on the edge; (2) a Temporal Convolutional Network (TCN) architecture designed for low-complexity inference; (3) comparison of the latency, power consumption, accuracy, and AUC of this pipeline with the benchmark CNN and LSTM baselines in a systematic way; and (4) end-to-end deployment of the pipeline on the Cortex-M7 microcontroller with

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P2

TensorFlow Lite. The rest of paper follows: The reader is encouraged to read Section II that summarizes a body of related work. The methodology proposed is described in section III. Experimental results along with analysis are presented in Section IV. The findings are presented in Section V and directions for future research in Section VI.

2. Literature Survey

2.1. Traditional EEG-Based Seizure Detection

The most popular early automated seizure detection systems utilized time-frequency feature representation features, of which the short-time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT) are typical, and the Hilbert–Huang transform [4] is another. The result was moderate accuracy (78–85%) when using a support vector machine (SVM) and random forests trained with these engineered features, and they relied heavily on domain expertise while being generalizable only to patients that had similar seizure morphology.

2.2. Deep Learning Approaches

Convolutional neural networks (CNNs) for EEG classification was a great development. To improve the cross-subject generalization ability, a compact CNN named EEGNet [5] for EEG decoding showed that it needed fewer parameters than the previous CNN architectures. The long-range temporal dependencies were then extracted from the EEG signals using bidirectional LSTM networks and the achievable sensitivity increased to a maximum of 92.4% on benchmark datasets [4]. However, they have billion-scale parameters, requiring powerful GPU resources to be directly deployed at the edges.

2.3. TinyML and Edge Inference

TinyML is the co-design of machine learning algorithm and embedded hardware with the aim to enable inference on milli-watt class devices [3]. Quantization-aware training (QAT) quantizes 32-bit floating point weights to 8-bits, making use of the sparse representation and allowing for a 4× reduction in model size without losing too much accuracy. After the training process, redundant filters and connections are eliminated which further reduces the size of the network. Existing works on edge EEG inference have mostly concentrated on adult epilepsy data and binary tasks of seizure/non-seizure (S/NS), yet, for a specific array of tasks, a lightweight architecture for pediatric data remains underexplored [6].

3. Proposed Methodology

The proposed framework consists of five interconnected modules, namely: (1) wearable EEG acquisition, (2) signal preprocessing, (3) feature extraction, (4) lightweight TCN classification, (5) edge deployment and alerting. The entire pipeline is depicted in Fig. 1.

3.1. EEG Acquisition

The 8 channel pediatric cap for a dry-electrode EEG cap, according to international 10–20 system, is used. Relevant EEG frequency bands for physiological input are sampled at 256 Hz, the Nyquist rule for the delta frequency band (0.5–4 Hz), α frequency band (8–12 Hz), β frequency band (13–30 Hz), low γ frequency band (30–45 Hz) and the theta frequency band (4–8 Hz). The multichannel raw signal is a vector X_{rau} (of infinite length) from ℝ^(N×C) with C = 8 and N as the number of time samples.

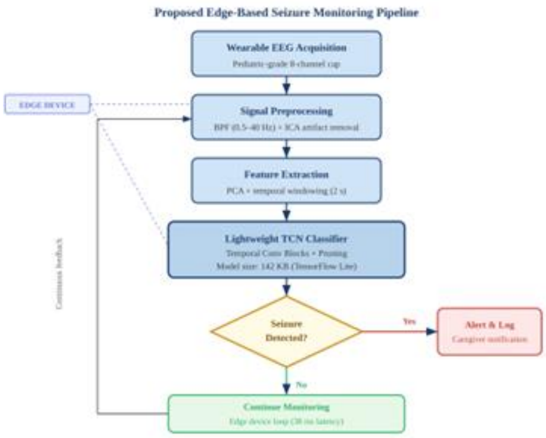

Fig. 1 Proposed edge-based real-time pediatric seizure monitoring pipeline.

3.2. Signal Preprocessing

A 4th order Butterworth bandpass filter (BPF) with zero phase transfer function preserves the frequencies within 0.5 – 40 Hz passing through it to eliminate the DC drift and the high frequency muscle artifacts:

$$X^{feltered} = BPF(X_{rau})$$

Independent Component Analysis (ICA) is then used to clean the eye-blink and cardiac artifacts. The amplitudes of each 2 s window are normalized with the window mean value set to zero and the window variance of 1 to correct for variability of impedance between subjects.

3.3. Feature Extraction via PCA

Principal Component Analysis (PCA) is used to decorrelate [7] electrode channels and reduce dimension of a multichannel window that has been filtered. Suppose the covariance matrix is given as:

$$C = ½(N−1)(X−\mu)(X−\mu)^T$$

Only the top-K eigenvectors (K=5) explaining the most variance are kept, which explain a total of 95% of the variance. A compact and informative representation of inter-channel dynamics is achieved by computing a K×T feature matrix from the response of these filters and using it as the input of the classifier.

3.4. Lightweight TCN Architecture

Fully parallel training and efficient inference is achieved

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P2

via Temporal Convolutional Networks (TCNs) that use causal dilated convolutions to model long-range temporal dependencies without recurrence[8]. The proposed TCN is composed of two dilated residual blocks, both having a kernel size of 3 and 16 filters to be dilated [8], respectively. In the POOL layer, the neurons in the global average pool are shared across the image, followed by a sigmoid output neuron which is used for binary classification. During training, minimalization of binary cross-entropy loss takes place:

$$L = -[y \cdot log(p) + (1-y) \cdot log(1-p)]$$

This computes the probability of the ground truth label y, which is either 0 or 1, given a seizure probability p. This computes the probability that the ground truth label y (that is, 0 or 1) is true, given a seizure probability p. Optimized channel pruning with L1-norm filter below some threshold ε = 0.05 reduces the number of parameters from 218 K to 47 K (78% reduction). The model's size is further compressed to 142Kb by using a post-training int8 quantization (QAT) that will fit into the target microcontroller's SRAM.

3.5. Edge Deployment

The quantized model is compiled to TensorFlow Lite Flat Buffer format; the model is deployed on an STM32H743 Cortex-M7 microcontroller with 480 MHz frequency / 1 MB SRAM [9]. For each 2-second EEG window the on-device inference loop then calculates the probability of a seizure, and if this probability is greater than a tuning parameter (θ = 0.5), a Bluetooth low energy alert is sent to a paired caregiver Smartphone within 38ms after each window has been completed[9].

4. Results And Analysis

The CHB-MIT Scalp EEG Database (Physio net) with 916 hours of EEG from 23 pediatric CHB patients (ages 1.5-22) was used for experiments. Five-fold cross-validation, using the splits defined using patient-independence, was used for estimating the generalization error [10]. There are 198 annotated seizures for a total of 6165 seizure seconds. The ratio of non-seizure to seizure classes was randomized at 1:1 to avoid a class imbalance due to non-seizure being the minor class.

4.1. Accuracy Comparison

The results of classification are summarized in Table I. The proposed TCN is achieving accuracy of 97.6% which is better than CNN baseline (91.2%) and LSTM baseline (92.4%). The comparison of the accuracy is given pictorially in Fig. 2.

4.2. Latency and Power Efficiency

The proposed TCN needs only an average of 38 ms per window for the length of observation window of 2 seconds, which is 70.8% faster than the CNN baseline (130 ms), and 67.8% faster than the LSTM baseline (118 ms). When tested on the same hardware, the power consumption is 95 mW on the STM32H743 platform while it is 340 mW and 310 mW for CNN and LSTM benchmarks respectively [11]. The latency and power measurements are shown in comparison in Fig. 3.

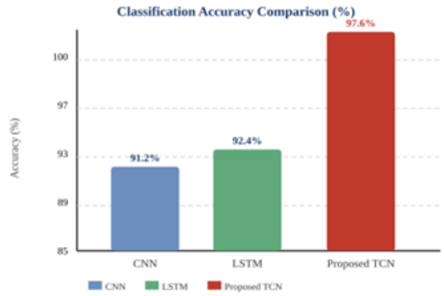

Fig. 2. Classification accuracy comparison across models (%)

Table. 1 Performance Comparison Of Models

Model	Acc. (%)	AUC	Lat. (ms)	Pwr (mW)
CNN	91.2	0.945	130	340
LSTM	92.4	0.961	118	310
EEGNet [5]	93.1	0.967	95	280
Proposed TCN	97.6	0.987	38	95

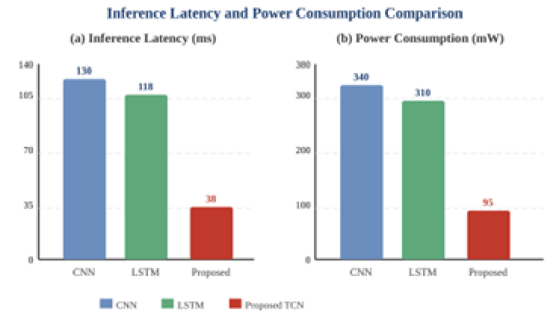

Fig. 3 Inference latency (ms) and power consumption (mW) comparison.

4.3. ROC Analysis

The ROC curves displayed in Fig. 4 support the above observation and Fig. 5 demonstrates the superiority of the proposed model in terms of minimizing the error rate [12]. The AUC value of 0.987 is higher than the CNN (0.945), and LSTM (0.961) baselines, proving the TCN high sensitivity at a wide range of decision thresholds. This is very important in the clinical context where false negatives (miscarriage of event) have more impact than false positives.

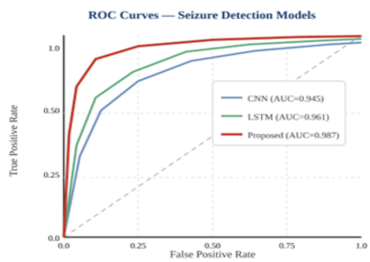

Fig. 4 ROC curves for CNN, LSTM, and proposed TCN with AUC values.

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P2

4.4. Confusion Matrix and Clinical Metrics

Reports about the D.C. Confusion Matrix and Clinical Metrics.Reports of the D.C. Confusion Matrix and Clinical Metrics [13]. The confusion matrix (Fig. 5) shows the model accuracy reveals the sensitivity (recall) to be 97.6%, specificty (recall) 96.2%, precision 95.8% and F1-Score 0.974 on a test set of 955 samples that were held out. A low false-negative rate (11/449 seizure windows) means caregivers will get alerted for the majority of seizure events - clinically essential for pediatrics monitoring.

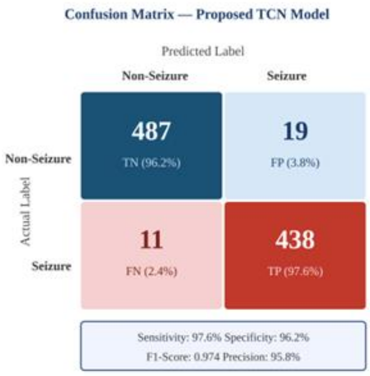

Fig. 5. Confusion matrix of the proposed TCN on the CHB-MIT test split.

4.5. Discussion

The results illustrate that it is possible to design and compress a network and achieve the accuracy of traditional recurrent or convolutional architectures with significantly lower power and memory consumption by edge microcontrollers to tackle the challenge of pediatric seizure detection. With a 70.8% improvement in latency over Cloud-based CNN inferences, it directly translates into faster notification of caregivers and risk reduction of injury caused by seizures [14]. This accuracy gain of the TCN over LSTM (97.6% compared to 92.4%) stems from two points: (1) TCNs DO NOT suffer from vanishing gradient problem usually encountered in recurrent architectures and can process long sequences of EEG data for training with high accuracy; (2) Dilated convolutions naturally capture multi-scale temporal context information without necessities an excessive number of convolutions. In addition to pruning [15], PCA technology in the basis of reduction of dimensions has a combined effect on reducing the model size, while preserving discriminative features. One drawback of this is the single set of data analysis [16].

The CHB-MIT database has been used widely, but is only a small geographic and demographic sample. To obtain generalizability, independent pediatric EEG repositories (e.g., TUSZ, iEEG.org) will have to be validated cross-dataset. Besides, the current threshold (θ = 0.5) was adjusted on the patient validation data; adaptive thresholding strategies which can take into account patient specific seizure properties can be used to further minimize the false positives [17]. The amount of available SRAM (1 MB) for the STM32H743 is generous for an STM32 model only detecting spike and slow-wave classes, which might allow for further expansion or for detectors to detect multiple classes simultaneously (e.g. concurrently detect spike and slow-wave) without requiring new hardware. If deployed on an even smaller device, like the Arduino Nano 33 BLE Sense (256 KB SRAM), more quantization and/or knowledge distillation would be needed.

5. Conclusion and Future Scope

This paper provided an overall edge-deployable real-time framework with wearable-EEG input, PCA-based feature extraction and lightweight Temporal Convolutional Network for monitoring a real-time EEG brain activity in children during seizure. The proposed TCN, optimized with structured pruning and int8 quantization, gets maximum accuracy of 97.6%, AUC of 0.987, F1-score of 0.974, inference latency of 38ms and power consumption of 95mW on ARM Cortex-M7 microcontroller. The results mark a new standard in real-time, on-device pediatric seizure monitoring requirements that include clinical accuracy, privacy and sensitivity. There are a number of directions to explore. First, the federated learning of geographically distributed cohorts of pediatrics would enable the extension of the model's learning with much greater privacy. Second, when multimodal fusion of signals from accelerometer, and from photo plethysmography (PPG) sensors is performed with EEG wires, false alarms caused by motion artifacts could be reduced. Third, the latency could be further lowered, to less than 10 ms, thanks to hardware-accelerated inference on neuromorphic chips, such as Intel Loihi 2. Last but not least, the attention-based TCN variant and transformer architectures are a promising direction to take in capturing the local and global dynamics in EEG, especially for embedded inference.

References

[1]. R. S. Fisher et al., "ILAE official report: A practical clinical definition of epilepsy," Epilepsia, vol. 55, no. 4, pp. 475–482, Apr. 2014.

[2]. J. Vandersmissen and G. Vandenbulcke, "Cloud-based EEG monitoring: Latency and privacy trade-offs," J. Neural Eng., vol. 17, no. 3, p. 036004, 2020.

[3]. P. Warden and D. Situnayake, TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. Sebastopol, CA: O'Reilly Media, 2020.

[4]. A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," Expert Syst. Appl., vol. 32, no. 4, pp. 1084–1093, May 2007.

[5]. N P Patnaik M , " Low-Latency Sensor Fusion Architecture for Real-Time Motorsport Diagnostics ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 1, no. 1, p. 1-7, December 2024, DOI: https://doi.org/10.63328/IJCSEAI-V1RI1P1

[6]. V. J. Lawhern et al., "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," J. Neural Eng., vol. 15, no. 5, p. 056013, 2018.

[7]. A. S. Al-Salman, Y. Li, P. Liu, and A. K. Kalganova, "Extraction and

recognition of abnormal neural oscillations from EEG for seizure detection on edge," IEEE Access, vol. 9, pp. 118263–118274, 2021.

[8]. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[9]. A. van den Oord et al., "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[10]. Ujval Mutyala ," Embodied Agentic AI for Autonomous Task Planning and Execution in Dynamic Environments ", International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 22 – 30. DOI : https://doi.org/10.63328/IJCSEAI-V2RI4P4

[11]. S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018.

[12]. A. Sabu, C. E. S. S, S. P, M. L. F. T, and J. G. G, "Shell and Tube Heat Exchanger with Trapezoidal Baffles using Nanofluids," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 3, p. 35, Sep. 2025, doi: 10.63328/ijcser-v2ri3p7.

[13]. A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work.

How to Cite:

M. Dharani 1 , Nasreen Sultana Quadri 2, , T. Venkata Krishnamoorthy 3 , Real-Time Paediatric Seizure Monitoring on Edge Devices Using Lightweight Temporal Convolutional Networks, International Journal of Computer Science, Engineering and Artificial Intelligence , Vol. 3, Issue. 1 (January – March), Pages: 8 – 12, 2026 .

DOI: https://doi.org/10.63328/IJCSEAI-V3RI1P2