Journal of Edge Computing, 2026, Vol. **5**, Iss. 1, pp. 173–188 10.55056/jec.1172

Optimising seizure prediction with reduced computational resources using depthwise CNN

Ritesh Dhananjay Nikose, Suchismita Chinara

National Institute of Technology Rourkela, Jagda, Sector-1, Rourkela, 769008, India

Abstract. Existing deep learning models for epileptic seizure prediction are accurate but parameter-heavy, which limits their deployment on wearable and other resource-constrained edge devices. We present DSCNN_Net, a 3D depthwise separable convolutional network operating on Mel-frequency cepstral coefficient (MFCC) features extracted from scalp EEG. On the CHB-MIT dataset DSCNN_Net reaches 89.58% sensitivity with 11,714 parameters and 45.75 KB of weight memory – roughly an order of magnitude fewer parameters than comparable CNN baselines at similar sensitivity. Replacing standard 3D convolution with its depthwise separable form reduces the per-layer multiply–accumulate cost by approximately 10× without a loss of predictive performance, supporting real-time operation on low-power edge platforms.

Keywords: epileptic seizure prediction, deep learning models, EEG signals, computational efficiency in healthcare, energy consumption

1. Introduction

More than 50 million people worldwide live with neurological disorders, and epileptic seizures are among the most prevalent manifestations [15]. Seizures arise from abnormal electrical activity in the brain and are routinely characterised through electroencephalography (EEG), which records voltage fluctuations from scalp electrodes.

Raw scalp EEG is noisy, non-stationary and high-dimensional, so any seizure-prediction pipeline must combine careful preprocessing with a feature representation that exposes the slow, low-frequency activity associated with the preictal state. We use Mel-frequency cepstral coefficients (MFCCs) for this purpose; the details are given in section 3.2.

Predictors based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have driven most of the recent improvements in seizure prediction accuracy [16, 18, 35]. The price of these gains is parameter count and inference cost, which become limiting on the wearable and embedded hardware that is the natural home of a continuous-monitoring system. Inference latency, on-device memory and energy budget are all linear in (or worse than) the parameter count [20, 29, 31, 40]; sections 3.8 and 3.6 make this dependence explicit.

This paper presents DSCNN_Net, a depthwise separable 3D CNN that maintains competitive sensitivity on CHB-MIT while reducing parameter count to roughly one tenth of comparable baselines. The remainder of the paper reviews related work (section 2), describes the dataset, preprocessing and architecture (section 3), defines the evaluation protocol (section 4), and reports comparative results on accuracy, memory, inference time, training time and energy (section 5).

2. Literature review

This section reviews CNN-based seizure prediction, energy-efficient variants and hybrid CNN–recurrent architectures, focusing on the trade-off between reported sensitivity and parameter count.

0009-0009-7622-8611 (R. D. Nikose); 0000-0002-2766-7820 (S. Chinara)
919CS5011@nitrkl.ac.in (R. D. Nikose); suchismita@nitrkl.ac.in (S. Chinara)
https://www.nitrkl.ac.in/CS/~suchismita/ (S. Chinara)

Received	Accepted	Published	Version of record
2025-10-06	2026-05-19	2026-05-19	2026-05-21

2.1. Convolutional neural networks in seizure prediction

Convolutional neural networks (CNNs) are the dominant approach for feature extraction from raw EEG signals in seizure prediction. Khan et al. [16] used six CNN layers with wavelet-coefficient inputs, reporting 86.6–87.8% sensitivity with about 186,918 parameters. Truong et al. [33] used three CNN layers over short-time Fourier transform (STFT) spectrograms, achieving 81.2% sensitivity with 197,010 parameters. Ozcan and Erturk [24] applied a 3D CNN with an image-based EEG representation and reported 85.7% sensitivity using 156,441 parameters. Zhang et al. [42] combined two CNN layers with common spatial pattern features and a Butterworth band-pass filter, reaching 92.2% sensitivity with 194,420 parameters. Tian et al. [32] developed a multi-view feature-learning model with three 3D-CNN blocks and four fully connected layers, reaching 96.66% sensitivity at the cost of 2,880,000 parameters. These works establish CNN-based prediction as accurate but parameter-heavy.

2.2. Energy-efficient approaches

A second line of work targets energy and parameter efficiency. Zhao, Yang and Sawan [43] combined neural architecture search (NAS) with a CNN, obtaining 93.48% sensitivity at 68,038 parameters. Abdelhameed and Bayoumi [1] used a two-dimensional deep convolutional autoencoder with a bidirectional long short-term memory (Bi-LSTM) network, reaching 98.79% sensitivity at 139,600 parameters. Wang et al. [34] proposed a stacked 1D CNN for seizure onset detection on long-term scalp and intracranial EEG, reporting 88.14% segment-level sensitivity (99.54% segment-level accuracy) with 105,538 parameters. Qiu, Wang and Jiao [26] proposed LightSeizureNet, which stacks dilated 1D convolution, depthwise 1D convolution, global average pooling and a linear layer; it reports 96.49% sensitivity at 198,300 parameters.

2.3. Advanced architectures

Li et al. [19] used a fully convolutional nested LSTM for automatic seizure detection, reporting 95.42% sensitivity with 72,600 parameters. Such hybrid CNN–recurrent designs raise sensitivity but reintroduce parameter overhead from the recurrent layers.

2.4. Key insights from literature

Table 1 summarises the reported sensitivities and parameter counts of the methods above. Sensitivity above 90% is now routine, but most published architectures still use more than 50,000 parameters, which limits deployment on wearable or embedded hardware. The model proposed in this work targets the same accuracy regime with one order of magnitude fewer parameters by combining depthwise separable 3D convolution with MFCC inputs.

3. Materials and methods

3.1. Dataset

Acquiring annotated medical data is constrained by legal and privacy requirements, so this study uses the publicly available Children's Hospital Boston (CHB-MIT) scalp EEG database [14]. Table 2 summarises the full corpus.

The corpus comprises 23 cases recorded from 22 paediatric patients (five males aged 3–22 years and 17 females aged 1.5–19 years) with intractable seizures. Signals were captured from 22 scalp electrodes placed according to the international 10–20 system, sampled at 256 Hz with 16-bit resolution. Most files contain 23 EEG signals; some contain 24 or 26 [14].

For seizure prediction, interictal periods are defined as being at least 4 hours before seizure onset and 4 hours after seizure end. Seizures occurring less than 30 minutes apart are considered a single

Table 1
Summary of existing methods.

Paper	Dataset	Sensitivity (%)	Algorithm	Total parameters	Trainable parameters	Non-trainable parameters
Truong et al. [33]	CHB-MIT	81.2	CNN	197,010	196,786	244
Khan et al. [16]	CHB-MIT	87.8	CNN	186,918	186,918	0
Ozcan and Erturk [24]	CHB-MIT	85.7	3D CNN	156,441	156,441	0
Zhang et al. [42]	CHB-MIT	92.2	CNN	194,420	194,420	0
Zhao, Yang and Sawan [43]	CHB-MIT	93.48	CNN	68,038	68,038	0
Abdelhameed and Bayoumi [1]	CHB-MIT	98.79	Bi-LSTM	139,600	139,600	0
Wang et al. [34]	CHB-MIT	88.14	1D CNN	105,538	105,538	0
Li et al. [19]	CHB-MIT	95.42	CNN, LSTM	72,600	72,600	0
Qiu, Wang and Jiao [26]	CHB-MIT	96.49	1D CNN	198,300	198,300	0
Tian et al. [32]	CHB-MIT	96.66	3D CNN	2,880,000	2,880,000	0

Table 2
Full CHB-MIT scalp EEG corpus [14].

Dataset	EEG type	Patients	Channels	Seizures	Recording hours
Children's Hospital Boston-MIT	Scalp	22 (23 cases)	22	163	844

seizure, using the onset of the leading seizure for prediction purposes. Only patients with fewer than 10 seizures per day were considered, as predicting seizures for patients with frequent seizures (every 2 hours on average) is less critical. Consequently, data from 12 patients with at least three leading seizures and 3 hours of interictal recording were utilised as explained in [33].

Furthermore, only recordings containing the following channels were considered: ['FP1-F7', 'F7-T7', 'T7-P7', 'P7-O1', 'FP1-F3', 'F3-C3', 'C3-P3', 'P3-O1', 'FP2-F4', 'F4-C4', 'C4-P4', 'P4-O2', 'FP2-F8', 'T8-P8','F8-T8', 'P8-O2', 'FZ-CZ', 'CZ-PZ', 'P7-T7', 'T7-FT9', 'FT9-FT10', 'FT10-T8']. According to these criteria, only 12 patients are included in this work, as shown in table 3.

Table 3
CHB-MIT database used in this work.

Number of patients	Interictal hours	No. of seizures
pt1	17.7	7
pt2	23.1	3
pt3	21.9	6
pt5	14.4	5
pt9	50.0	4
pt10	26.0	6
pt14	4.2	5
pt18	25.0	6
pt19	23.1	3
pt20	20.8	5
pt21	21.6	4
pt23	14.2	5
Total	99.7	59

3.2. Data pre-processing

We compute Mel-frequency cepstral coefficients (MFCCs) over the raw EEG signal. MFCCs concentrate spectral information in the low-frequency band, which carries most of the discriminative content for seizure prediction [9, 13]; the exact MFCC pipeline used here is detailed in section 3.3.

The CHB-MIT recordings contain mains-frequency interference at 60 Hz. We remove this by band-stop filtering 57–63 Hz and 117–123 Hz (the second band suppresses the first harmonic).

Interictal segments greatly outnumber preictal segments in CHB-MIT, which biases a softmax classifier toward the majority class. We address this imbalance with the overlapped-window sampling scheme of Truong et al. [33]: additional preictal segments are generated by sliding a 30-second window with a subject-specific step S chosen so that the preictal and interictal counts match in the training split. The resulting balanced training set lets the network learn a discriminative preictal–interictal boundary [3, 5].

3.3. Signal processing

MFCCs are computed from the short-time Fourier transform (STFT) of each 30-second window. The i-th cepstral coefficient is

$$MFCC_i = \sum_{m=1}^{M} s(m) \cos\left[i \left(m - \tfrac{1}{2} \right) \frac{\pi}{M} \right], \quad i = 1, 2, \ldots, L, \tag{1}$$

where $s(m)$ is the log energy of the m-th Mel filter bank, M is the number of Mel bands and L is the cepstral order [13].

The pipeline (figure 1) is: window the EEG with a sliding Hamming window to limit spectral leakage, take the discrete Fourier transform (DFT) of each window [39], group the magnitude spectrum into Mel-scaled triangular filter banks, take the logarithm of the band energies, and apply a discrete cosine transform (DCT) to obtain the MFCCs [13, 22, 25, 30]. Concretely, EEG is sampled at 256 Hz and split into 30-second windows; a 1 Hz high-pass filter and the band-stop filters above are applied before feature extraction; a 256-point FFT with library-default parameters yields 20 MFCCs and 16 time frames per channel. With 22 EEG channels, each segment becomes a tensor of shape $1 \times 22 \times 20 \times 16$ (temporal block × channels × coefficients × frames).

3.4. Depthwise separable 3D convolution

The proposed architecture replaces the standard 3D convolution layer with its depthwise separable form, which factorises a 3D convolution into two cheaper stages (figure 2).

The depthwise stage convolves each input channel independently with its own 3D filter, sliding across the spatial and depth dimensions. This preserves the channel count and produces one feature map per channel, but each filter only sees a single channel, so the multiply count scales with the input–filter product rather than with the product of input channels and output channels.

The pointwise stage applies a bank of $1 \times 1 \times 1$ filters across all input channels at each position, mixing per-channel feature maps into the desired number of output channels [41]. The combination of the two stages has been shown to reduce parameter count by roughly an order of magnitude relative to a standard 3D convolution at matched receptive field [38].

In the EEG setting the input is naturally three-dimensional: scalp channels, MFCC coefficients and short-term time frames. A 1D CNN captures only temporal variation; a 2D CNN captures time–frequency patterns per channel independently. A 3D depthwise separable convolution captures inter-channel structure together with time–frequency correlations at a fraction of the cost of the standard 3D form.

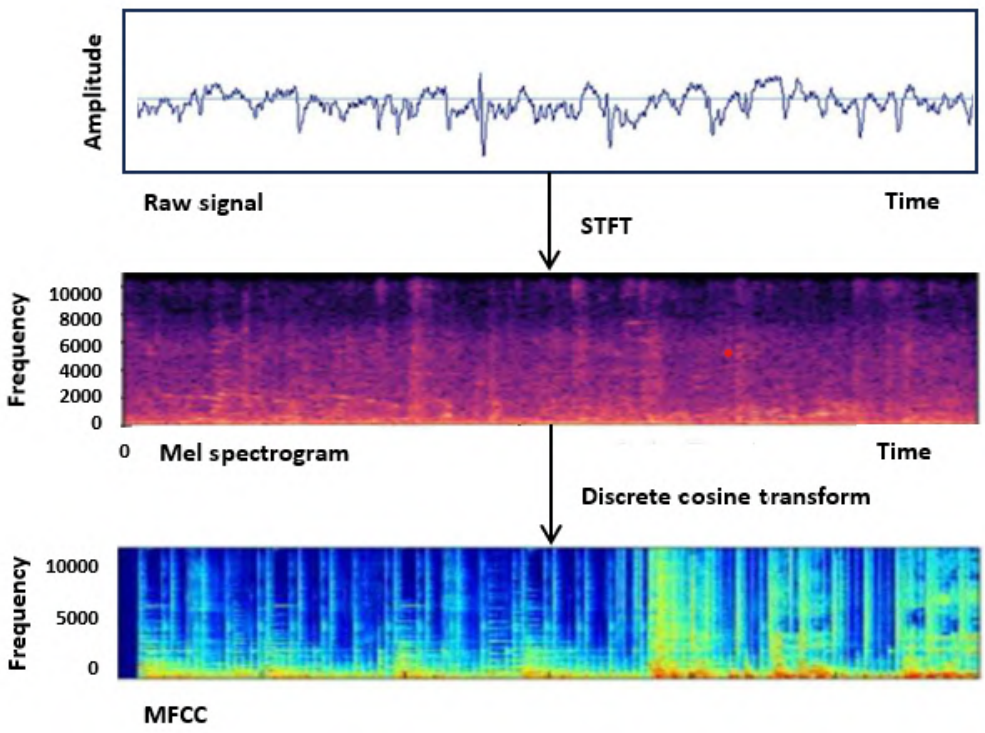

Figure 1: Signal conversion.

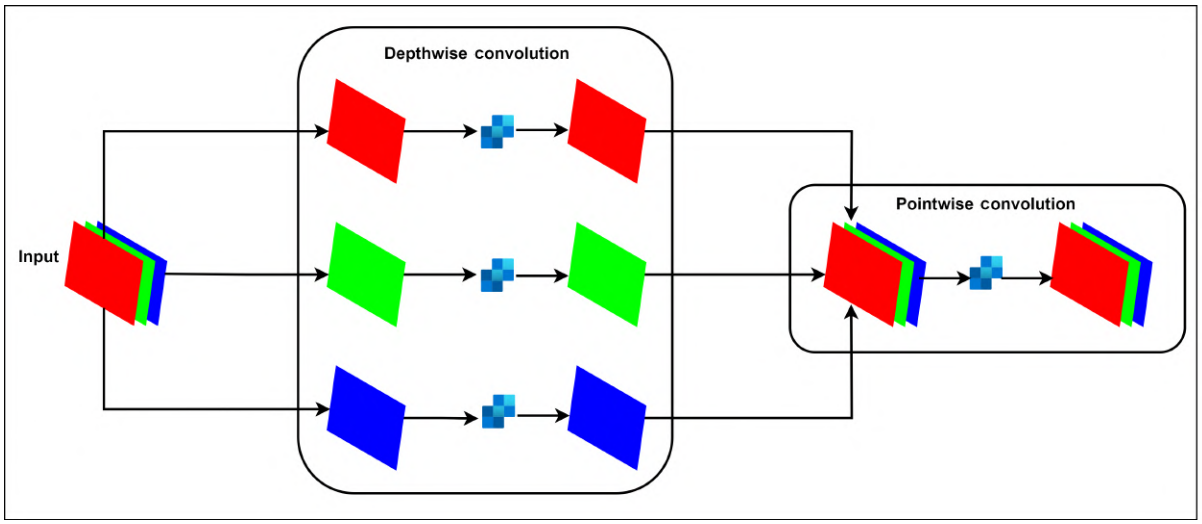

Figure 2: Depthwise separable 3D convolution: a depthwise stage applies one 3D filter per input channel, followed by a $1 \times 1 \times 1$ pointwise stage that mixes channels.

3.5. Cost comparison: standard 3D convolution vs. depthwise separable 3D convolution

The computational cost for a standard 3D convolution operation can be described as:

$$\text{Cost}_{\text{std\_3D}} = D_i \times H_i \times W_i \times C_i \times C_o \times K_d \times K_h \times K_w \tag{2}$$

where D_i, H_i, W_i, and C_i are the input depth, height, width, and channels; C_o is the output channels; K_d, K_h, K_w are the filter dimensions.

The total computational cost is the sum of the depthwise convolution and the pointwise convolution:

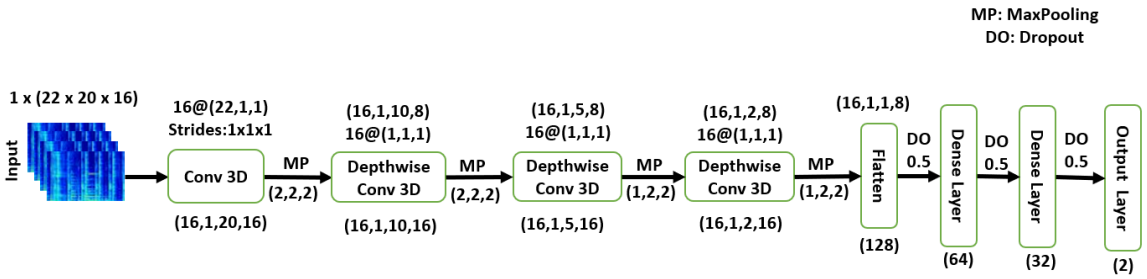

Figure 3: Model architecture.

$$\text{Cost}_{\text{DS\_3D}} = D_i \times H_i \times W_i \times C_i \times K_d \times K_h \times K_w + C_i \times C_o \times D_i \times H_i \times W_i \tag{3}$$

The ratio of the costs:

$$\frac{\text{Cost}_{\text{DS\_3D}}}{\text{Cost}_{\text{std\_3D}}} = \frac{1}{C_o} + \frac{1}{K_d \times K_h \times K_w} \tag{4}$$

The output-channel count C_o is typically 32, 64 or 128, so $1/C_o$ is small. The factor $1/(K_d K_h K_w)$ is dominated by the kernel size: even for a kernel as small as $2\times2\times2$ this term is $1/8 \approx 0.125$. Their sum is therefore approximately 0.1 in practice, giving the depthwise separable form an order-of-magnitude advantage in parameter count over the standard 3D convolution.

3.6. Proposed model architecture: DSCNN_Net

The proposed model, DSCNN_Net, comprises an eight-layer architecture, depicted in figure 3. The initial layer is a 3D convolutional layer, followed by 3D max pooling with a pool size of (2, 2, 2) and batch normalisation. This layer uses a stride of $1 \times 1 \times 1$ and a kernel size of (2, 2, 1), with the ReLU activation function. The subsequent three layers are depthwise separable 3D convolutions with a depth multiplier of 2, kernel size of $1 \times 1 \times 1$, and stride of $1 \times 1 \times 1$, followed by max pooling with sizes $2 \times 2 \times 2$, $1 \times 2 \times 2$, and $1 \times 2 \times 2$, respectively. These are followed by a flattened layer with a dropout rate of 0.5, and two dense layers with sigmoid activation functions, each containing 64 and 32 neurons, respectively, with a dropout rate of 0.5. The final output layer consists of two neurons with a softmax activation function. All convolutional and dense layers used the default Keras Glorot uniform weight initialisation, with zero-bias initialisation, and a fixed random seed of 42 for reproducibility.

L_2 weight decay of 10^{-4} is applied to every depthwise separable 3D convolution. The network is trained with categorical cross-entropy and the Adam optimiser at a learning rate of 5×10^{-5}.

Training uses a custom early-stopping callback that monitors the validation loss after each epoch and terminates training once the loss falls below 10^{-5}. Each model is trained for at most 1000 epochs with batch size 1 on a single NVIDIA T4 GPU (16 GB GDDR6); the small batch size is dictated by the length of the temporal EEG sequences.

The total computational cost of DSCNN_Net is obtained by summing the costs of its layers:

1. First layer 3D convolution:

$$\text{Cost}_{\text{std\_3D}} = D_i \times H_i \times W_i \times C_i \times C_o \times K_d \times K_h \times K_w \tag{5}$$

2. Subsequent separable depthwise 3D convolution layers:

$$\text{Cost}_{\text{DS\_3D}} = D_i \times H_i \times W_i \times C_i \times K_d \times K_h \times K_w + C_i \times C_o \times D_i \times H_i \times W_i \tag{6}$$

For L separable depthwise convolution layers, the total cost is:

$$\text{Total\_Cost}_{\text{DS\_3D}} = L \times (D_i \times H_i \times W_i \times C_i \times K_d \times K_h \times K_w + C_i \times C_o \times D_i \times H_i \times W_i) \quad (7)$$

3. Each dense layer with N neurons and input size I has a cost:

$$\text{Cost}_{\text{Dense}} = I \times N \quad (8)$$

4. Total computational cost:

$$\text{Total Cost} = \text{Cost}_{\text{Conv3D}} + \text{Total\_Cost}_{\text{DS\_3D}} + \sum_{i=1}^{2} \text{Cost}_{\text{Dense},i} \quad (9)$$

The model is implemented in TensorFlow 2.4.1 with Python 3.7.2. The data are split patient-specifically: for each patient, 20% of the segments are held out for testing and 80% are used for training/validation. Early stopping on validation loss and a fixed 1000-epoch upper bound guard against overfitting on this small dataset.

To assess the practical feasibility of edge deployment, the trained model is additionally executed on a Raspberry Pi 4 running Ubuntu. Inference latency is measured by feeding 100 random input tensors of the same shape as a real segment; instantaneous voltage and current are recorded with a KWS-V20 USB tester (3–9 V, 0–3 A) to estimate device-level power consumption.

3.7. Post-processing

Some patients produce isolated false positives during testing. We suppress them with a k-of-n rule: an alarm is raised only if at least k of the most recent n window-level predictions are positive. The values $k = 8$, $n = 10$ gave the best trade-off between sensitivity and false-alarm rate on the validation split and are used in all reported results.

3.8. Estimation of energy consumption

Energy efficiency is a primary constraint when deploying deep learning models on Internet-of-Things (IoT) and wearable devices, and it has been studied extensively in both the systems and the machine learning literature [7, 8, 12, 17, 27]. To a first approximation, the energy consumed by a deep learning model is dominated by the number of trainable parameters that must be moved between memory and compute units.

A naïve estimate of inference energy counts only the multiply–accumulate (MAC) operations performed by the model. However, on modern accelerators MAC arithmetic accounts for only a small share of the total energy budget; data movement – in particular, fetching parameters from DRAM into the on-chip caches and feature-map traffic between layers – dominates [10, 36]. Because each parameter incurs at least one DRAM access during a forward pass, the parameter count is a useful proxy for inference energy:

$$E_{\text{model}} = N_{\text{params}} \cdot E_{\text{data}}, \quad (10)$$

where N_{params} is the parameter count and E_{data} is the system-dependent energy cost of moving a single parameter through the memory hierarchy. Although E_{data} is hard to pin down on a given platform, it is approximately constant across models on a fixed hardware target, so reducing N_{params} is the most direct way to reduce inference energy [11]. The comparative analysis in section 5 therefore reports both per-inference GPU energy and parameter count as complementary metrics.

4. System evaluation

We adopt the seizure prediction horizon (SPH) and seizure occurrence period (SOP) of Maiwald et al. [21]: the SOP is the window during which a seizure is expected; the SPH is the interval between the alarm and the start of the SOP (figure 4). A prediction is correct if a seizure occurs after the SPH and

within the SOP (figure 5); an alarm without a seizure inside the SOP is a false alarm (figure 6). Bou Assi et al. [6] call the SPH the *intervention time*: it must be long enough for precautionary action but the SOP must not be so long that patients are left in a state of constant alarm. Following Truong et al. [33] we use SPH = 5 min and SOP = 30 min.

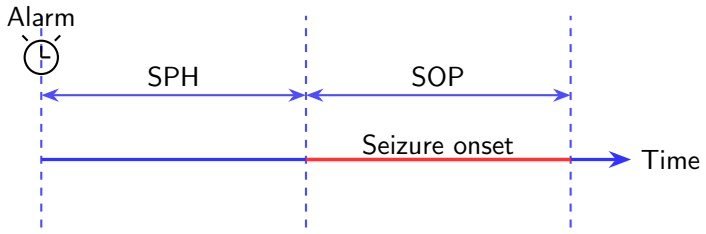

Figure 4: Definition of SPH and SOP.

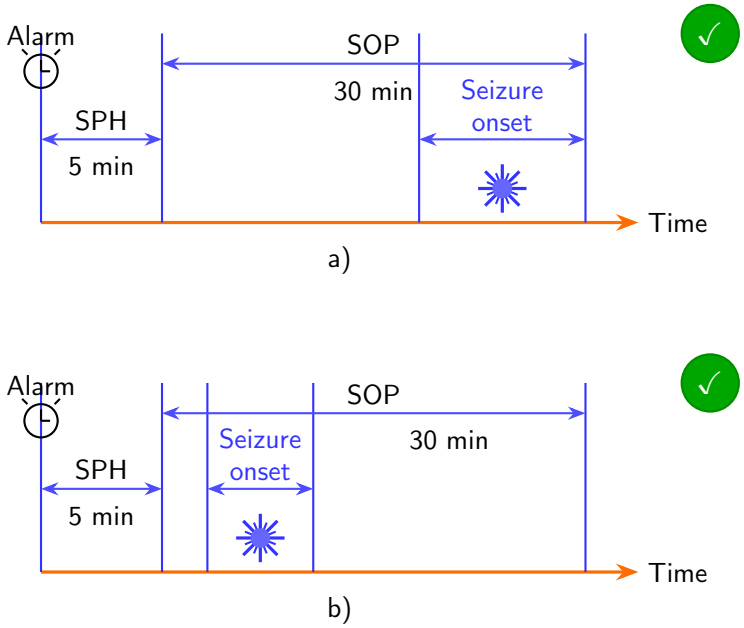

Figure 5: Accurate prediction: a) as seizure ends just before the end of SOP; b) as seizure onset within the SOP.

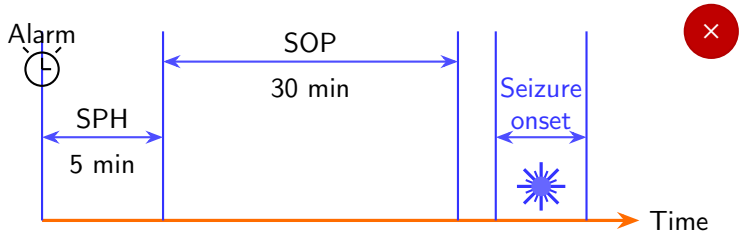

Figure 6: False alarm.

The predictive performance of the proposed approach is evaluated by comparing it against a random predictor. To assess the reliability of predictions, the false positive rate (FPR) is utilised, representing the probability of raising an alarm in the SOP, as proposed by [28]. The probability of independently predicting at least n out of N seizures is calculated using the binomial equation:

$$P(X \geq n) = \sum_{k=n}^{N} \binom{N}{k} p^k (1-p)^{N-k},$$

where $P(X \geq n)$ is the cumulative probability of predicting at least n seizures, $\binom{N}{k}$ is the binomial coefficient, p is the false positive rate (FPR), and N is the total number of seizures. For each patient, the p-value is determined based on the patient's average FPR and the number of seizures (n) successfully predicted by the proposed method. A p-value threshold of less than 0.05 indicates that the proposed approach outperforms the random predictor with statistical significance. The p-values for each patient are shown in table 4. For Patient 02 ($p = 0.83$) and Patient 09 ($p = 0.06$) the model does not outperform random prediction at the 0.05 significance level. Predictability depends on the strength and consistency of the preictal signature, which varies between subjects; the overall performance therefore reflects an average across patients rather than a per-subject guarantee.

Table 4
Per-patient accuracy and statistical significance against the random predictor. The seizure count reflects events evaluated after applying the leading-seizure and preictal-availability criteria of section 3.2; it is lower than the 59 leading seizures listed in table 3 because events without a complete preictal window were excluded from evaluation.

Patient	Seizures evaluated	Accuracy (%)	p-value
01	7	89.81	<0.01
02	3	84.00	0.83
03	5	100.00	0.01
05	5	99.23	<0.01
09	4	82.08	0.06
10	6	91.82	0.01
14	4	100.00	<0.01
18	3	100.00	<0.01
19	2	100.00	<0.01
20	5	51.46	<0.01
21	4	90.20	<0.01
23	5	86.33	<0.01
Total / mean	53	89.58	–

Following Bates, Hastie and Tibshirani [4], we use cross-validation for robust evaluation: the data are split into N folds, with $N-1$ folds for training and the remaining fold for validation, repeated N times so each data point is used for validation exactly once [2]. The proposed model is additionally tested on 20% of the original data, held out from training, as in Mormann et al. [23].

5. Results and comparative analysis

This section compares DSCNN_Net with state-of-the-art baselines on sensitivity, memory footprint, inference time, GPU energy and training time. Not every baseline reports every metric, so the set of comparators varies between figures; in each case we include only those for which the metric is published.

5.1. Sensitivity

Per-patient sensitivities are reported in table 4; the mean is 89.58%. As figure 7 shows, this sits within the range spanned by the CNN baselines (81.2–93.48%). DSCNN_Net is below the highest reported sensitivity (Zhao, Yang and Sawan [43], 93.48%) but is competitive while using approximately 5× to 20× fewer parameters; the design target is parameter efficiency under similar sensitivity rather than maximum sensitivity in isolation.

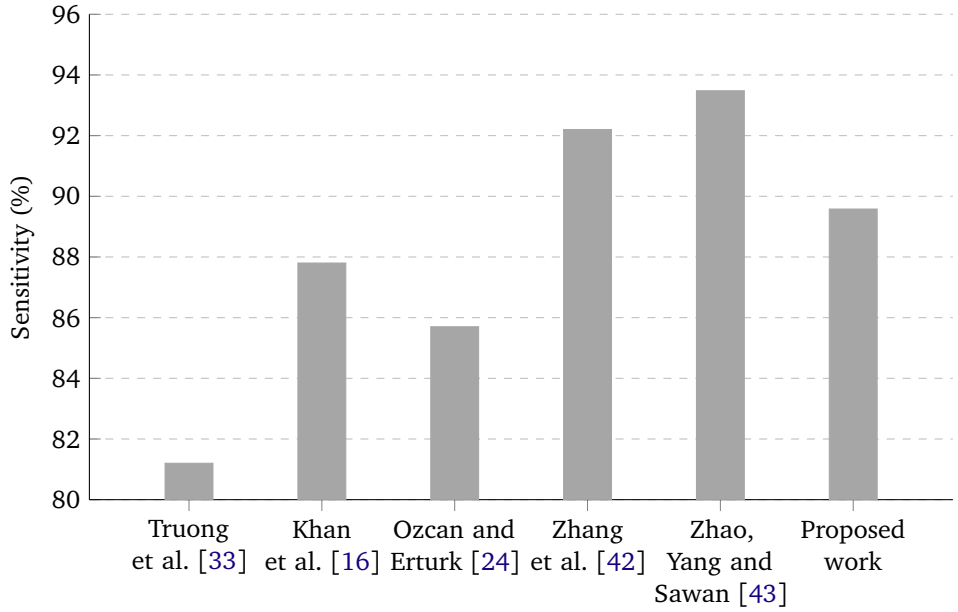

Figure 7: Reported sensitivity of state-of-the-art seizure-prediction models on CHB-MIT.

5.2. Memory consumption

DSCNN_Net needs 45.75 KB of weight memory (figure 8), against 611–770 KB for the four full-precision CNN baselines. Zhao, Yang and Sawan [43] report 256.77 KB without quantisation and 45.22 KB with 16-bit quantisation; quantisation closes the gap on storage but reintroduces extra MAC operations and dequantisation overhead at inference time. The relevant constraint on a wearable device is not aggregate storage but memory bandwidth and cache footprint: a smaller weight tensor needs fewer DRAM fetches per forward pass and therefore reduces both energy per inference and steady-state battery drain.

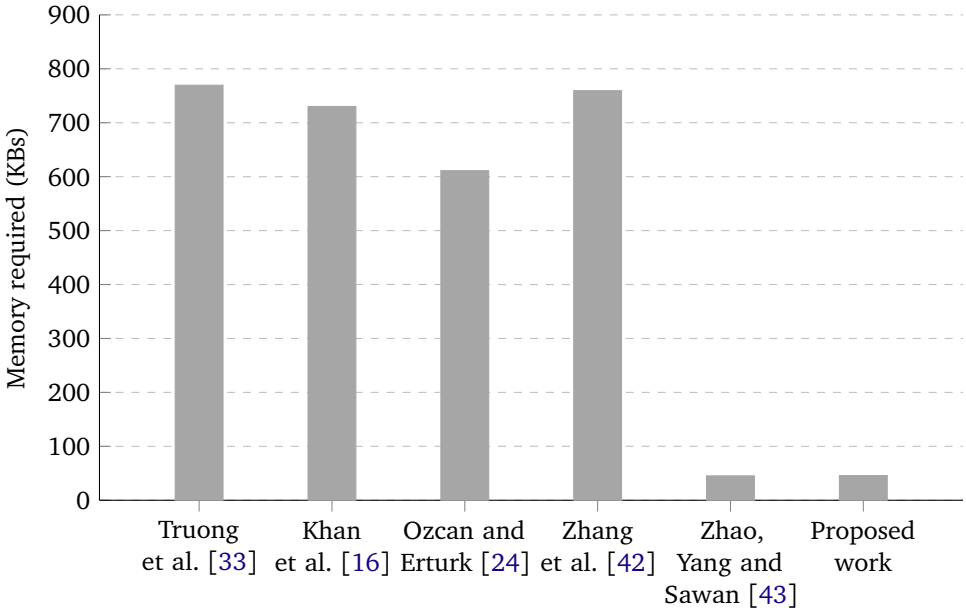

Figure 8: Weight memory of the compared models.

Table 5
Detailed comparison with the state-of-the-art methods.

Paper	Sensitivity (%)	Algorithm	Total parameters	Training time (s)	Inference time (s)	Memory (KB)	Energy (J)
Truong et al. [33]	81.20	CNN	197,010	5.1312	0.001573	769.57	0.0498
Khan et al. [16]	87.80	CNN	186,918	2.5078	0.000431	730.15	0.0136
Ozcan and Erturk [24]	85.70	3D CNN	156,441	2.0021	0.000480	611.10	0.0152
Zhang et al. [42]	92.20	CNN	194,420	1.3897	0.000326	759.45	0.0103
Zhao, Yang and Sawan [43]	93.48	CNN	68,038	–	–	45.22	–
DSCNN_Net (this work)	89.58	Depthwise 3D CNN	11,714	1.7230	0.000320	45.75	0.0101

5.3. Inference time

Inference latency was measured on an NVIDIA T4 GPU as the mean over four runs of 100 random input tensors per model (figure 9). DSCNN_Net needs 0.000320 s per sample, against 0.000431 s for Khan et al. [16] and 0.000480 s for Ozcan and Erturk [24].

On a Raspberry Pi 4 (Ubuntu), the per-sample latency for DSCNN_Net rises to 0.226102 s. This is two to three orders of magnitude slower than the T4 measurement but is the more realistic figure for the wearable deployment scenario, and remains well below the 5-minute SPH used in this work.

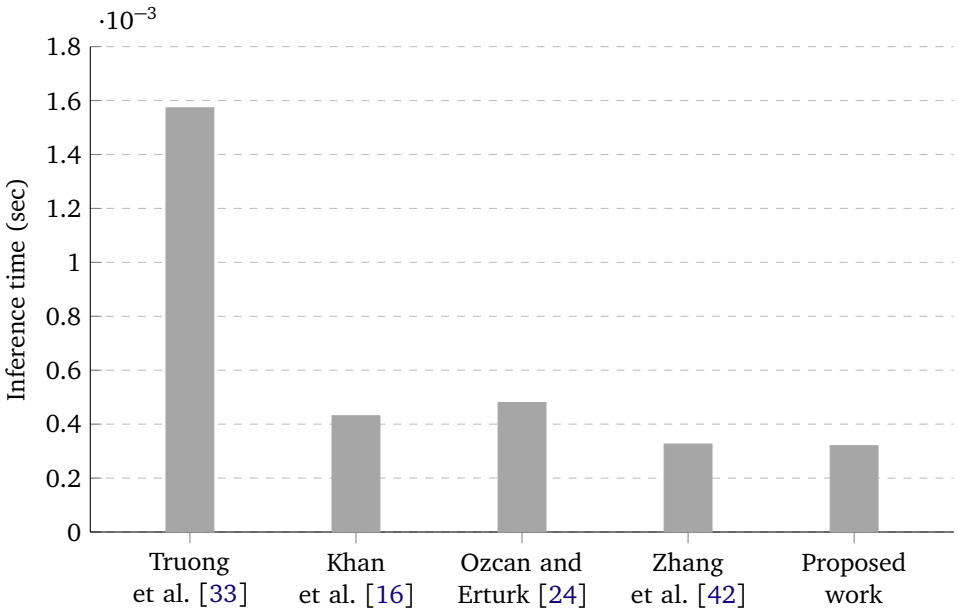

Figure 9: Mean per-sample inference time on the T4 GPU.

5.4. Power consumption

Per-inference GPU energy was measured on an NVIDIA T4 (Google Colab) using `nvidia-smi`, integrating power over the inference window for each model. Results are reported in table 5. These figures reflect only GPU-die power and exclude on-board memory traffic across the GPU's memory hierarchy, which can be a substantial fraction of the total system energy budget on modern accelerators [37].

Because GPU-die power is a partial picture, parameter count remains the more robust proxy for inference energy on memory-bound hardware (section 3.8). DSCNN_Net uses an order of magnitude fewer parameters than the four full-precision CNN baselines (figure 10), which is what drives the

lower measured energy per inference in table 5.

To corroborate the GPU measurement on real edge hardware, inference power was also measured on a Raspberry Pi 4 (Ubuntu) using a KWS-V20 USB tester. The device drew approximately 5.15–5.16 V at 0.65–0.67 A during inference, giving an average power $P = V \cdot I \approx 3.4$ W. With the measured 0.226102 s per inference, this gives $E = P \cdot t \approx 0.77$ J per prediction – a sustainable energy budget for a battery-powered wearable.

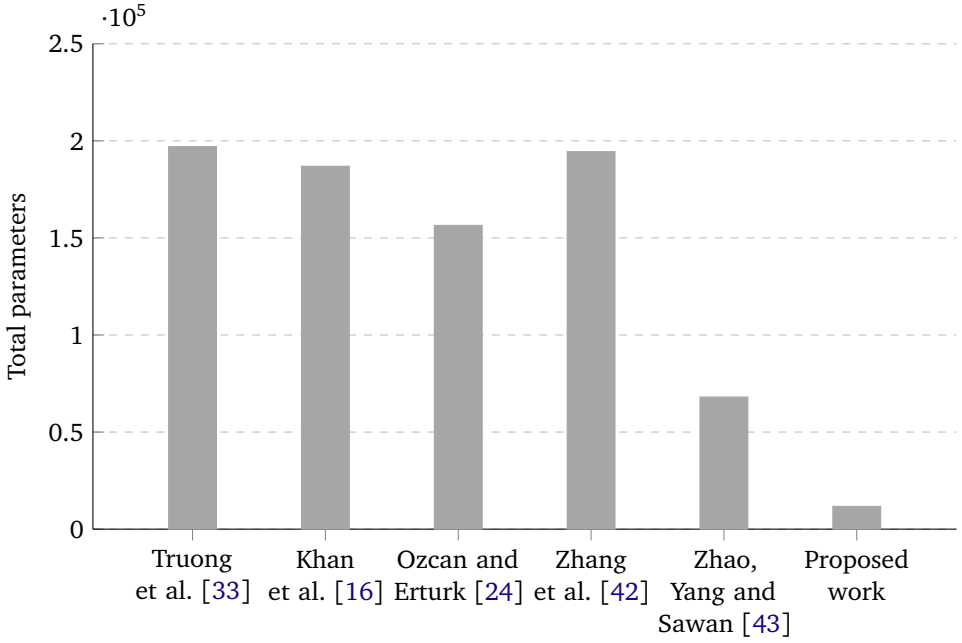

Figure 10: Total parameter counts of the compared models.

5.5. Training time

A shorter training cycle matters in settings that require frequent retraining, such as patient-specific re-personalisation. For each model we trained over 100 epochs on 100 random input segments and recorded the mean per-epoch time; results are in figure 11. DSCNN_Net trains in 1.7230 s/epoch, against 2.5078 s for Khan et al. [16] and 2.0021 s for Ozcan and Erturk [24].

5.6. Comprehensive comparative analysis

Table 5 consolidates the per-metric comparisons of the preceding subsections. DSCNN_Net is comparable in sensitivity to the four full-precision CNN baselines, while using approximately one order of magnitude fewer parameters and around 13× less weight memory. Inference latency, training time and GPU energy per inference are all lower than or comparable to the baselines, so the parameter saving does not come at the cost of throughput.

6. Limitations

This work is evaluated only on the CHB-MIT corpus, so generalisation across acquisition devices, channel montages and clinical populations is not yet established. Energy efficiency is reported from GPU power telemetry and a single Raspberry Pi 4 measurement; deployment on a dedicated low-power inference accelerator may show a different power profile. Predictive performance varies between patients – for Patient 02 and Patient 09 the model does not statistically outperform the random predictor – and the headline sensitivity is therefore a mean over subjects rather than a per-subject guarantee.

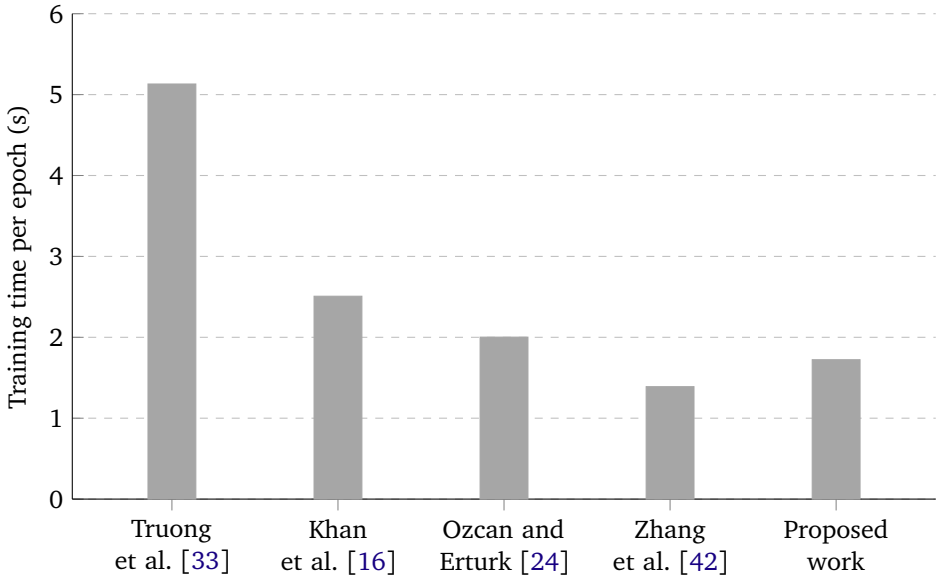

Figure 11: Mean training time per epoch on the T4 GPU.

The current results do not include an ablation: the individual contributions of MFCC preprocessing, depthwise separable convolution and the dense head have not been isolated. Future work will add per-component ablations, evaluate quantisation as a second axis of memory and energy reduction, and validate the model on additional EEG corpora and on dedicated edge accelerators.

7. Conclusion

DSCNN_Net is a depthwise separable 3D CNN for seizure prediction from scalp EEG. On CHB-MIT it reaches 89.58% mean sensitivity with 11,714 parameters – roughly one tenth of the parameters of comparable CNN baselines at similar sensitivity – and uses 45.75 KB of weight memory. GPU energy per inference is the lowest among the compared models, and the Raspberry Pi 4 deployment confirms that the architecture is small enough to run on commodity edge hardware within the chosen 5-minute prediction horizon. The result supports the broader argument that, in this regime, parameter efficiency is at least as actionable as raw accuracy for moving seizure-prediction systems out of the laboratory and onto wearables.

Funding

This research received no external funding.

Data availability statement

The research was conducted using the publicly available CHB-MIT Scalp EEG Database, which contains de-identified data to ensure the privacy and confidentiality of individuals. The implementation code and preprocessing scripts are available upon reasonable request to ensure full reproducibility of the reported results.

Conflicts of interest

The authors declare no conflict of interest.

Declaration on Generative AI

The authors have not employed any generative AI tools.

References

[1] Abdelhameed, A. and Bayoumi, M., 2021. A Deep Learning Approach for Automatic Seizure Detection in Children With Epilepsy. *Frontiers in Computational Neuroscience*, 15, p.650050. Available from: https://doi.org/10.3389/fncom.2021.650050.

[2] Alotaiby, T.N., Alshebeili, S.A., Alotaibi, F.M. and Alrshoud, S.R., 2017. Epileptic Seizure Prediction Using CSP and LDA for Scalp EEG Signals. *Computational Intelligence and Neuroscience*, 2017(1), p.1240323. Available from: https://doi.org/10.1155/2017/1240323.

[3] Bandarabadi, M., Rasekhi, J., Teixeira, C.A., Karami, M.R. and Dourado, A., 2015. On the proper selection of preictal period for seizure prediction. *Epilepsy & Behavior*, 46, pp.158–166. Available from: https://doi.org/10.1016/j.yebeh.2015.03.010.

[4] Bates, S., Hastie, T. and Tibshirani, R., 2024. Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, 119(546), pp.1434–1445. Available from: https://doi.org/10.1080/01621459.2023.2197686.

[5] Binder, D.K. and Haut, S.R., 2013. Toward new paradigms of seizure detection. *Epilepsy & Behavior*, 26(3), pp.247–252. Available from: https://doi.org/10.1016/j.yebeh.2012.10.027.

[6] Bou Assi, E., Nguyen, D.K., Rihana, S. and Sawan, M., 2017. Towards accurate prediction of epileptic seizures: A review. *Biomedical Signal Processing and Control*, 34, pp.144–157. Available from: https://doi.org/10.1016/j.bspc.2017.02.001.

[7] Cai, E., Juan, D.C., Stamoulis, D. and Marculescu, D., 2017. Neuralpower: Predict and deploy energy-efficient convolutional neural networks. In: M.L. Zhang and Y.K. Noh, eds. *Proceedings of the Ninth Asian Conference on Machine Learning*. Yonsei University, Seoul, Republic of Korea: PMLR, *Proceedings of machine learning research*, vol. 77, pp.622–637. Available from: http://proceedings.mlr.press/v77/cai17a/cai17a.pdf.

[8] Chen, Y.H., Krishna, T., Emer, J.S. and Sze, V., 2017. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, 52(1), pp.127–138. Available from: https://doi.org/10.1109/JSSC.2016.2616357.

[9] Dissanayake, T., Fernando, T., Denman, S., Sridharan, S. and Fookes, C., 2020. Patient-independent Epileptic Seizure Prediction using Deep Learning Models. *CoRR*, abs/2011.09581. Available from: https://doi.org/10.48550/arXiv.2011.09581.

[10] Eranian, S., 2006. Perfmon2: a flexible performance monitoring interface for Linux. *Proceeding of the 2006 Ottawa Linux Symposium*. pp.269–288. Available from: https://www.kernel.org/doc/ols/2006/ols2006v1-pages-269-288.pdf.

[11] García-Martín, E., Rodrigues, C.F., Riley, G. and Grahn, H., 2019. Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134, pp.75–88. Available from: https://doi.org/10.1016/j.jpdc.2019.07.007.

[12] Gauen, K., Rangan, R., Mohan, A., Lu, Y.H., Liu, W. and Berg, A.C., 2017. Low-power image recognition challenge. *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*. pp.99–104. Available from: https://doi.org/10.1109/ASPDAC.2017.7858303.

[13] Guo, S. and Zhang, F., 2022. A SPCNN Model for Patient-Independent Prediction of Epilepsy Using MFCC Features. *2022 12th International Conference on Information Science and Technology (ICIST)*. pp.68–73. Available from: https://doi.org/10.1109/ICIST55546.2022.9926793.

[14] Guttag, J., 2010. CHB-MIT Scalp EEG Database. *PhysioNet*. Version 1.0.0. Available from: https://doi.org/10.13026/C2K01R.

[15] Hekim, M., 2012. ANN-based classification of EEG signals using the average power based on rectangle approximation window. *Electrical Review*, 88(8). Available from: https://archiwum.pe.org.pl/articles/2012/8/55.pdf.

[16] Khan, H., Marcuse, L., Fields, M., Swann, K. and Yener, B., 2018. Focal Onset Seizure Prediction

Using Convolutional Networks. *IEEE Transactions on Biomedical Engineering*, 65(9), pp.2109–2118. Available from: https://doi.org/10.1109/TBME.2017.2785401.

[17] Lane, N.D., Bhattacharya, S., Georgiev, P., Forlivesi, C. and Kawsar, F., 2015. An Early Resource Characterization of Deep Learning on Wearables, Smartphones and Internet-of-Things Devices. *Proceedings of the 2015 International Workshop on Internet of Things towards Applications*. New York, NY, USA: Association for Computing Machinery, IoT-App '15, pp.7–12. Available from: https://doi.org/10.1145/2820975.2820980.

[18] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P. and Lance, B.J., 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5), p.056013. Available from: https://doi.org/10.1088/1741-2552/aace8c.

[19] Li, Y., Yu, Z., Chen, Y., Yang, C., Li, Y., Allen Li, X. and Li, B., 2020. Automatic Seizure Detection using Fully Convolutional Nested LSTM. *International Journal of Neural Systems*, 30(04), p.2050019. Available from: https://doi.org/10.1142/S0129065720500197.

[20] Liu, X. and Richardson, A.G., 2021. Edge deep learning for neural implants: a case study of seizure detection and prediction. *Journal of Neural Engineering*, 18(4), p.046034. Available from: https://doi.org/10.1088/1741-2552/abf473.

[21] Maiwald, T., Winterhalder, M., Aschenbrenner-Scheibe, R., Voss, H.U., Schulze-Bonhage, A. and Timmer, J., 2004. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. *Physica D: Nonlinear Phenomena*, 194(3), pp.357–368. Available from: https://doi.org/10.1016/j.physd.2004.02.013.

[22] Mendonça, F., Mostafa, S.S., Ravelo-García, A.G., Morgado-Dias, F. and Penzel, T., 2019. A Review of Obstructive Sleep Apnea Detection Approaches. *IEEE Journal of Biomedical and Health Informatics*, 23(2), pp.825–837. Available from: https://doi.org/10.1109/JBHI.2018.2823265.

[23] Mormann, F., Andrzejak, R.G., Elger, C.E. and Lehnertz, K., 2007. Seizure prediction: the long and winding road. *Brain*, 130(2), pp.314–333. Available from: https://doi.org/10.1093/brain/awl241.

[24] Ozcan, A.R. and Erturk, S., 2019. Seizure Prediction in Scalp EEG Using 3D Convolutional Neural Networks With an Image-Based Approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(11), pp.2284–2293. Available from: https://doi.org/10.1109/TNSRE.2019.2943707.

[25] Phanikrishna, B.V., Prakash, A.J. and Suchismitha, C., 2023. Deep Review of Machine Learning Techniques on Detection of Drowsiness Using EEG Signal. *IETE Journal of Research*, 69(6), pp.3104–3119. Available from: https://doi.org/10.1080/03772063.2021.1913070.

[26] Qiu, S., Wang, W. and Jiao, H., 2023. LightSeizureNet: A Lightweight Deep Learning Model for Real-Time Epileptic Seizure Detection. *IEEE Journal of Biomedical and Health Informatics*, 27(4), pp.1845–1856. Available from: https://doi.org/10.1109/JBHI.2022.3223970.

[27] Rodrigues, C., Riley, G. and Luján, M., 2018. SyNERGY: An energy measurement and prediction framework for Convolutional Neural Networks on Jetson TX1. *PDPTA'18 - The 24th International Conference on Parallel and Distributed Processing Techniques and Applications*. Available from: https://research.manchester.ac.uk/en/publications/synergy-an-energy-measurement-and-prediction-framework-for-convol/.

[28] Schelter, B., Winterhalder, M., Maiwald, T., Brandt, A., Schad, A., Schulze-Bonhage, A. and Timmer, J., 2006. Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(1), p.013108. Available from: https://doi.org/10.1063/1.2137623.

[29] Shoaran, M., Haghi, B.A., Taghavi, M., Farivar, M. and Emami-Neyestanak, A., 2018. Energy-Efficient Classification for Resource-Constrained Biomedical Applications. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(4), pp.693–707. Available from: https://doi.org/10.1109/JETCAS.2018.2844733.

[30] Swain, M., Routray, A. and Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1), pp.93–120.

Available from: https://doi.org/10.1007/s10772-018-9491-z.

[31] Tan, T. and Cao, G., 2021. Deep Learning Video Analytics on Edge Computing Devices. *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. pp.1–9. Available from: https://doi.org/10.1109/SECON52354.2021.9491614.

[32] Tian, X., Deng, Z., Ying, W., Choi, K.S., Wu, D., Qin, B., Wang, J., Shen, H. and Wang, S., 2019. Deep Multi-View Feature Learning for EEG-Based Epileptic Seizure Detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(10), pp.1962–1972. Available from: https://doi.org/10.1109/TNSRE.2019.2940485.

[33] Truong, N.D., Nguyen, A.D., Kuhlmann, L., Bonyadi, M.R., Yang, J., Ippolito, S. and Kavehei, O., 2018. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*, 105, pp.104–111. Available from: https://doi.org/10.1016/j.neunet.2018.04.018.

[34] Wang, X., Wang, X., Liu, W., Chang, Z., Kärkkäinen, T. and Cong, F., 2021. One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG. *Neurocomputing*, 459, pp.212–222. Available from: https://doi.org/10.1016/j.neucom.2021.06.048.

[35] Xu, Y., Yang, J. and Sawan, M., 2022. Trends and challenges of processing measurements from wearable devices intended for epileptic seizure prediction. *Journal of Signal Processing Systems*, 94(6), pp.527–542. Available from: https://doi.org/10.1007/s11265-021-01659-x.

[36] Yang, T.J., Chen, Y.H., Emer, J. and Sze, V., 2017. A method to estimate the energy consumption of deep neural networks. *2017 51st Asilomar Conference on Signals, Systems, and Computers*. pp.1916–1920. Available from: https://doi.org/10.1109/ACSSC.2017.8335698.

[37] Yang, Z., Adamek, K. and Armour, W., 2024. Accurate and Convenient Energy Measurements for GPUs: A Detailed Study of NVIDIA GPU's Built-In Power Sensor. *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, pp.1–17. Available from: https://doi.org/10.1109/SC41406.2024.00028.

[38] Ye, R., Liu, F. and Zhang, L., 2019. 3D Depthwise Convolution: Reducing Model Parameters in 3D Vision Tasks. In: M.J. Meurs and F. Rudzicz, eds. *Advances in Artificial Intelligence*. Springer, *Lecture Notes in Computer Science*, vol. 11489, pp.186–199. Available from: https://doi.org/10.1007/978-3-030-18305-9_15.

[39] Zabidi, A., Mansor, W., Lee, Y.K. and Che Wan Fadzal, C.W.N.F., 2012. Short-time Fourier Transform analysis of EEG signal generated during imagined writing. *2012 International Conference on System Engineering and Technology (ICSET)*. pp.1–4. Available from: https://doi.org/10.1109/ICSEngT.2012.6339284.

[40] Zanetti, R., Arza, A., Aminifar, A. and Atienza, D., 2022. Real-Time EEG-Based Cognitive Workload Monitoring on Wearable Devices. *IEEE Transactions on Biomedical Engineering*, 69(1), pp.265–277. Available from: https://doi.org/10.1109/TBME.2021.3092206.

[41] Zhang, P., Lo, E. and Lu, B., 2020. High Performance Depthwise and Pointwise Convolutions on Mobile Devices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), pp.6795–6802. Available from: https://doi.org/10.1609/aaai.v34i04.6159.

[42] Zhang, Y., Guo, Y., Yang, P., Chen, W. and Lo, B., 2020. Epilepsy Seizure Prediction on EEG Using Common Spatial Pattern and Convolutional Neural Network. *IEEE Journal of Biomedical and Health Informatics*, 24(2), pp.465–474. Available from: https://doi.org/10.1109/JBHI.2019.2933046.

[43] Zhao, S., Yang, J. and Sawan, M., 2022. Energy-Efficient Neural Network for Epileptic Seizure Prediction. *IEEE Transactions on Biomedical Engineering*, 69(1), pp.401–411. Available from: https://doi.org/10.1109/TBME.2021.3095848.