Multi-Channel Vision Transformer for Epileptic Seizure Prediction

M1 NGUYEN ANH QUAN

Ramy Hussein, Soojin Lee, Rabab Ward

Biomedicines 2022, 10(7), 1551

https://doi.org/10.3390/biomedicines10071551

Reasons for Selecting the paper

- My research is related to the performance improvement of Vision Transformer Models for small-scale training data sets
- This paper provides a method to train The Vision Transformer architecture-based model with multichannel EEG data

Introduction

- Epilepsy is characterized by recurrent seizures that strike without warning.
- Seizure prediction has great potential to warn patients of an impending seizure so that they can take precautions to avoid any possible injury and administer rapid-acting medications.
- Currently, the electroencephalogram (EEG) is the most commonly used tool in seizure detection and prediction studies.

Introduction

- EEG activity of patients with epilepsy includes four prime states : preictal (right before seizure), ictal (seizure), postictal (immediately after seizure), and interictal (a seizure-free time period between the postictal and the preictal of consecutive seizures)
- The hand-crafted features (time domain features, frequency domain features, time-frequency domain features, and non-linear features) failed to attain clinical applicability due to a lack of generalization capacity.

→ A novel transformer-based algorithm (Vision Transformer) that accurately and robustly classifies preictal and interictal EEG activities has been proposed

Datasets

Dataset	CHB–MIT Scalp EEG Dataset[1]	Kaggle/American Epilepsy Society (AES) Invasive EEG Dataset[2]	Kaggle/Melbourne University Invasive EEG Dataset[3]
EEG data type	Scalp EEG	Invasive EEG	Invasive EEG
The number of subjects	22 people	2 people and 5 dogs	3 people
Sampling Frequency	256 Hz	400 Hz	400 Hz
The number of channels	23 channels	16 channels	16 channels
Measurement time	9-42 hours/person	7-12 months(5 dogs) 71.3 hours(female,70years old) 158.5 hours(female,48 years old)	559 days(female,22years old) 393 days(female, 51 years old) 374(female, 50 years old)

Datasets

- For 2 invasive EEG data sets:
  - Data were organized into 10-min EEG clips labeled “preictal” for pre-seizure data and “interictal” for inter-seizure
  - Preictal EEG data clips : EEG data for one hour before seizure with a five-minute offset
  - Interictal EEG data clips : EEG data were chosen randomly from the full EEG recordings

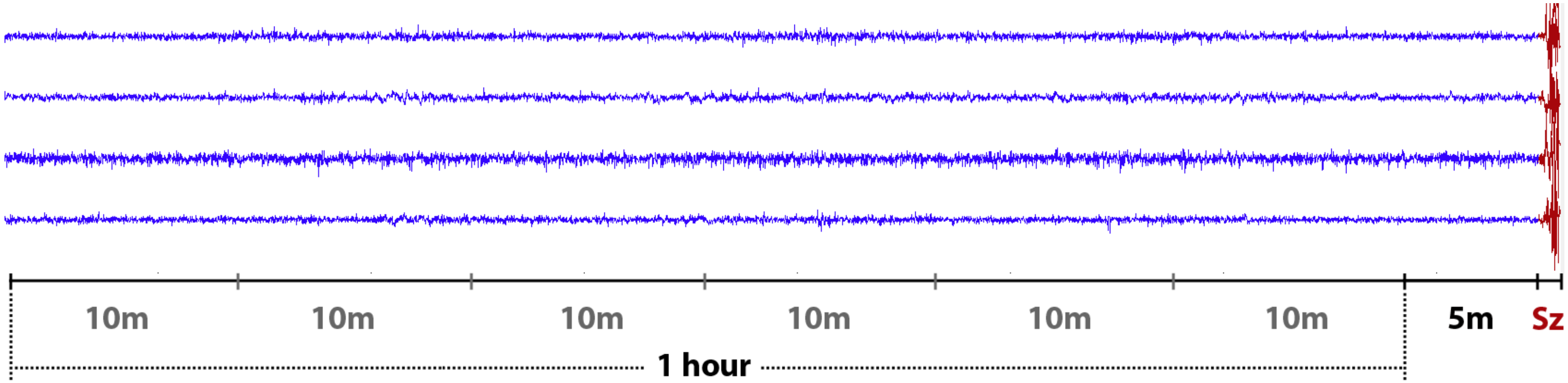


Figure 1. Examples of one-hour preictal (pre-seizure) EEG signals with a 5-min offset before seizures; Sz denotes the seizure onset. For convenience, only four channels are plotted.

Methodology

- Multi-channel Vision Transformer (MViT) is a variant of the original Vision Transformer (ViT)[4]
- The architecture consists of many different branches operating simultaneously on different EEG channels
- Before the EEG data is fed into the MViT, it is extracted the tempo-spectral feature at the pre-processing stage

EEG Pre-Processing

- Consists of 2 main procedures:
  - EEG Segmentation: Split each 10-min EEG clip into 10-sec EEG segments → 60 non-overlapping segments
  - Mapping EEG Segments into Images: Turning the results of EEG segmentation into image-like representations (scalogram) using continuous wavelet transform (CWT)

EEG Pre-Processing

Mapping 10-s EEG Segment in the invasive dataset into Scalogram

- The EEG segment is 10[sec] long and Sampling Frequency $f_s = 400[\mathrm{Hz}]$
  → The number of data-points $d = 10[s] \times 400[\mathrm{Hz}] = 4000$
- CWT is used to generate EEG power spectrum in the 3D domain
- 3D-to-2D projection (Proj) is used to produce the 2D time-frequency representations of EEG named "scalogram"

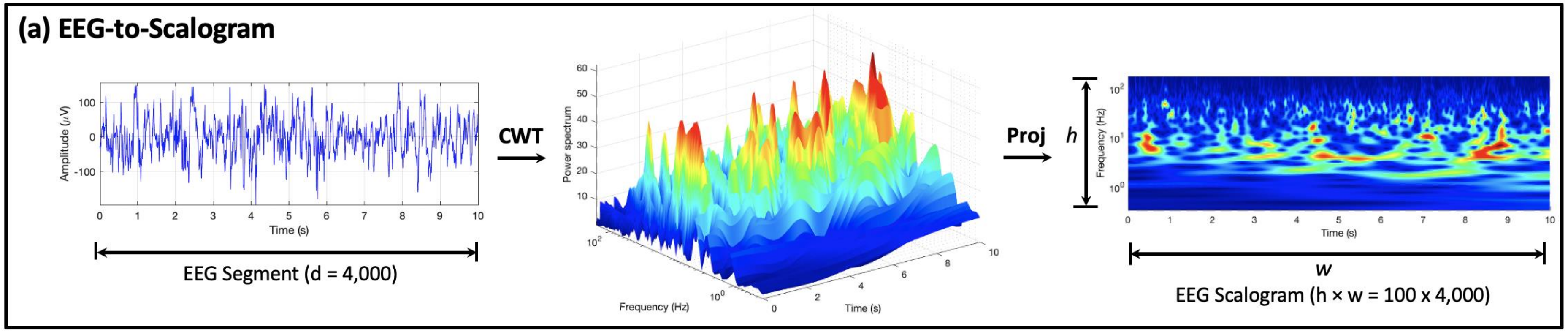


Figure 2. (a) EEG-to-scalogram conversion procedure

EEG Pre-Processing

Entire EEG preprocessing procedure:

- With a 10-min EEG data clip and N channels (Eg. N=16)
  → Data Shape is (Channels × Time[s] × Sampling Freq[Hz]) = (16×600×400)
- After segmenting into 60 segments with 10 seconds each
  → Data Shape is (Segments × Channels × data-points) = (60×16×4000)
- After wavelet transform for 60 segments
  → Data Shape is (Segments × Channels × Height × Width) = (60×16×100×4000)

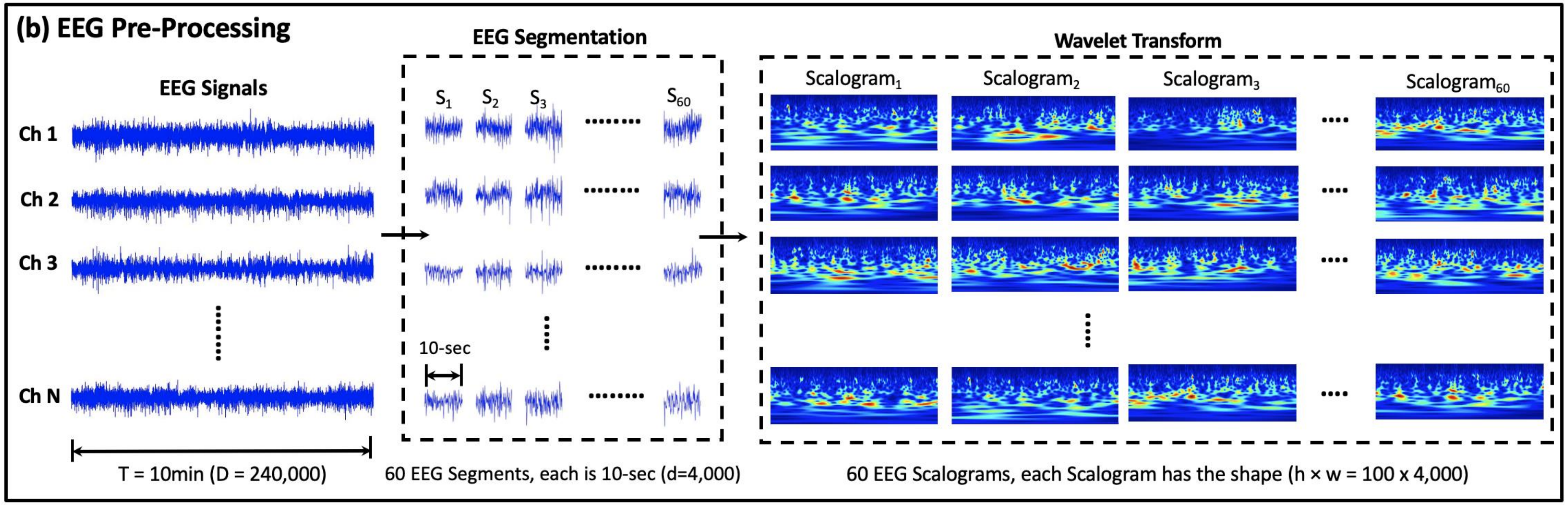


Figure 2. (b) EEG pre-processing approach

MViT for EEG Representation Learning

Assume P = 100, N = 16

→The Number of patches $L = \frac{H \times W}{P^2} = \frac{100 \times 4000}{100^2} = 40$

Lower-dimension D = 768 (*)

(L,D)=(40,768)

$x_p \in \mathbb{R}^{L \times P^2}$→ (L,P,P)=(40,100,100)

$x \in \mathbb{R}^{H \times W}$→ (H,W)=(100,4000)

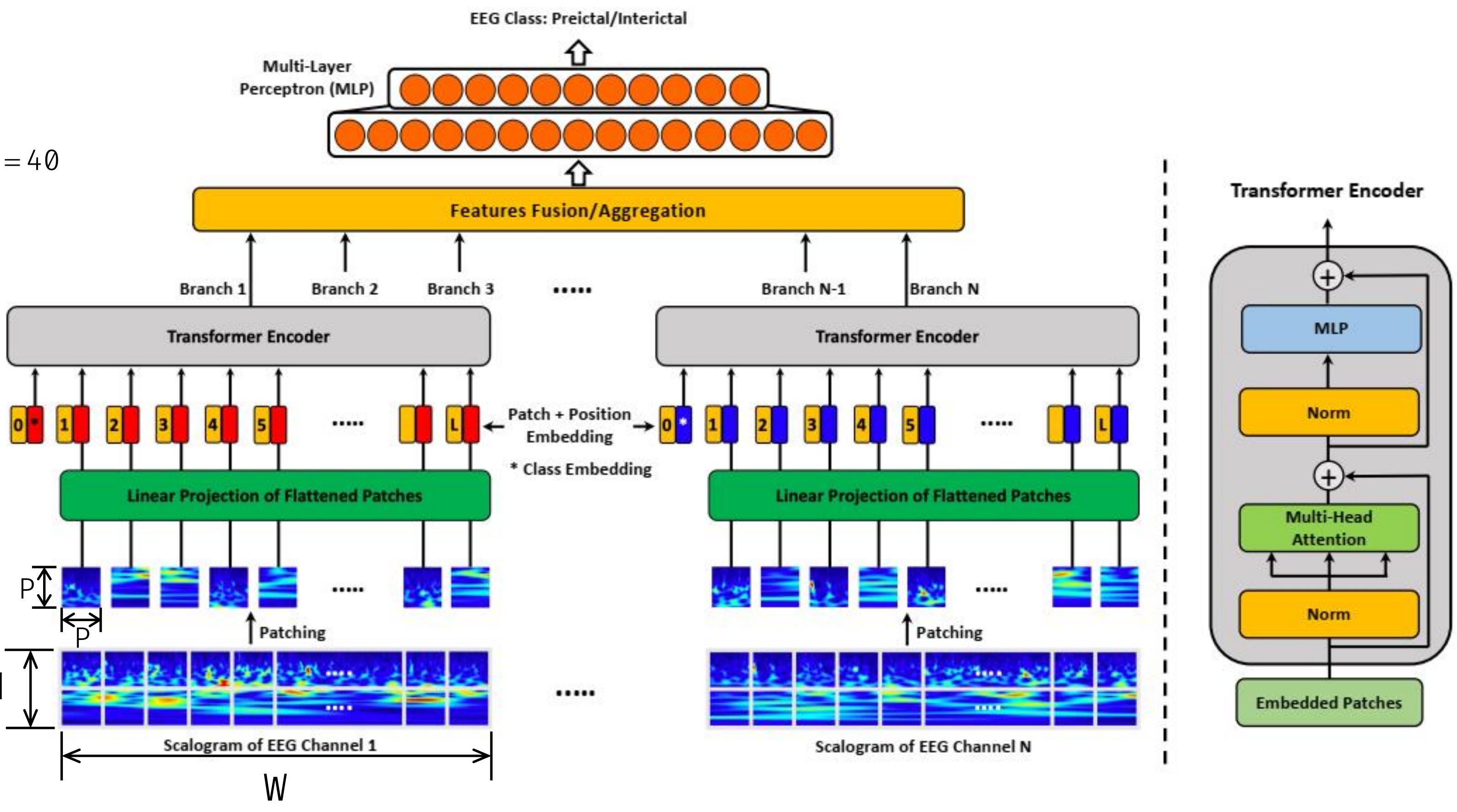


Figure 3. Framework of MViT for multi-channel EEG feature learning

(*) Hidden size of the ViT-Base model with 12 layers in the original ViT paper[4]

Self-Attention(Scaled Dot-Product Attention)

$$Q = X \times W^Q$$
$$K = X \times W^K$$
$$V = X \times W^V$$

Where, Q,K,V: Query,Key,Value matrix
X: Input matrix
W^Q ,W^K ,W^V: The parameter matrix that the model trained

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

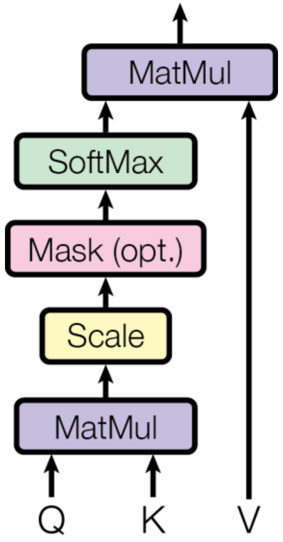
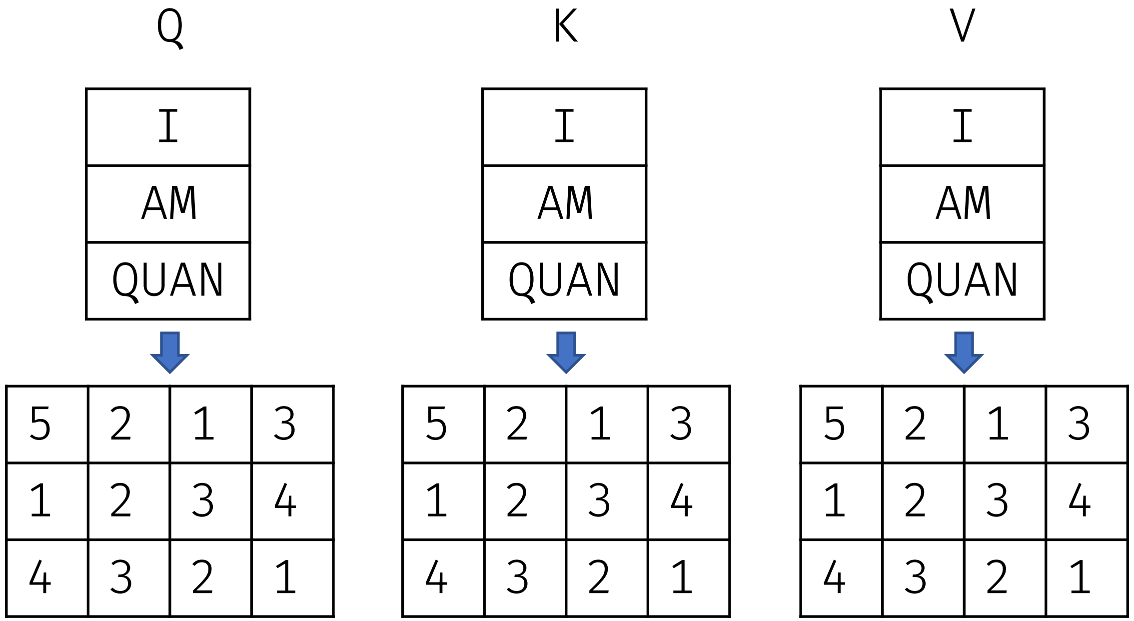


Figure 4. Scaled Dot-Product Attention[5]



Self-Attention(Scaled Dot-Product Attention)

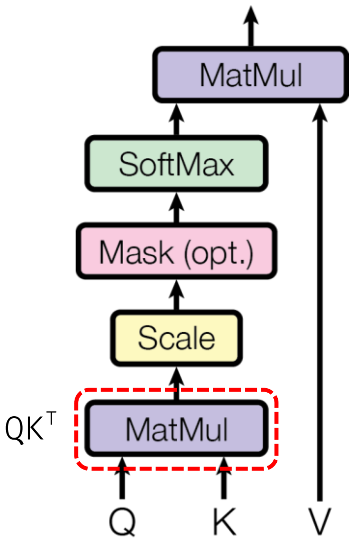


Figure 4. Scaled Dot-Product Attention[5]

Q

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

K

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

QK^T

K / Q	I	AM	QUAN
I	39	24	31
AM	24	30	20
QUAN	31	20	30

Self-Attention(Scaled Dot-Product Attention)

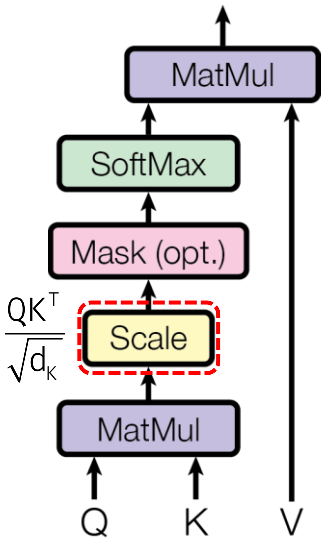
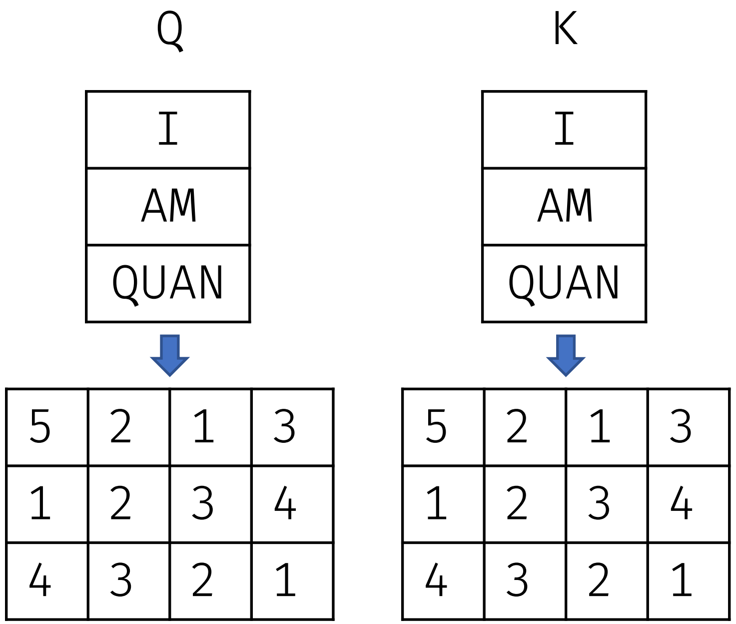


Figure 4. Scaled Dot-Product Attention[5]



$d_K = 3 \times 4 = 12$

$$\frac{QK^T}{\sqrt{d_K}}$$

Q \ K	I	AM	QUAN
I	11.3	6.9	8.9
AM	6.9	8.7	5.8
QUAN	8.9	5.8	8.7

Self-Attention(Scaled Dot-Product Attention)

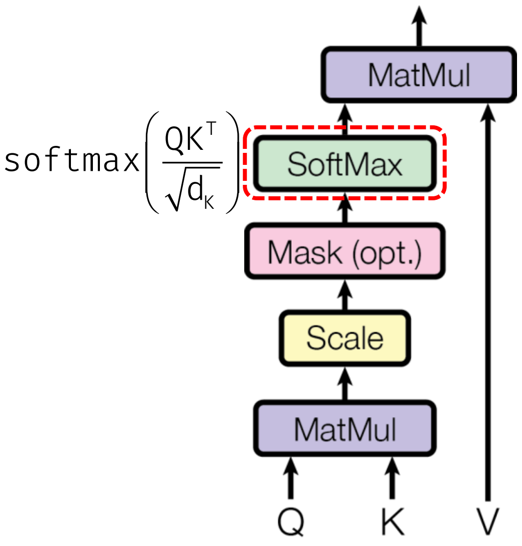


Figure 4. Scaled Dot-Product Attention[5]

Q

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

K

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

softmax $\left(\frac{QK^T}{\sqrt{d_k}}\right)$

Q \ K	I	AM	QUAN
I	0.9	0.01	0.09
AM	0.14	0.81	0.05
QUAN	0.56	0.02	0.42

sum equals 1

Self-Attention(Scaled Dot-Product Attention)

$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

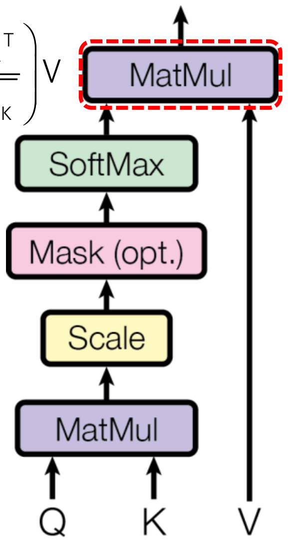


Figure 4. Scaled Dot-Product Attention[5]

Q

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

K

I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

V

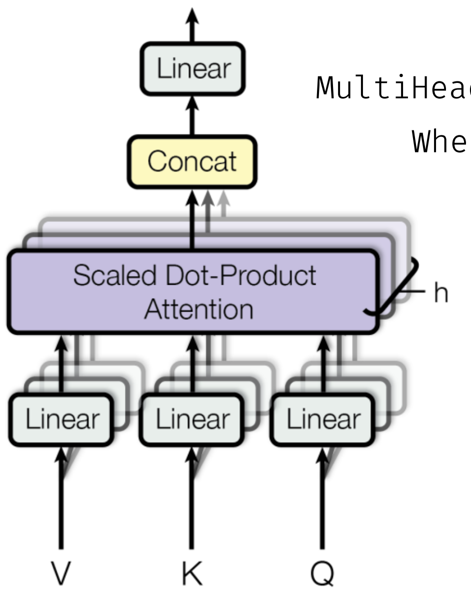
I
AM
QUAN

5	2	1	3
1	2	3	4
4	3	2	1

$\text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$

I	4.86	2.09	1.11	2.83
AM	1.71	2.05	2.67	3.72
QUAN	4.49	2.42	1.46	2.19

Multi-Head Attention



$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(head_1, ...head_h)W^O$$
$$\text{Where } head_i = \mathrm{Attention}(Q_i, K_i, V_i)$$

Figure 5. Multi-Head Attention consists of several attention layers[5]

× Problem: A word's attention will always "pay attention" to itself

Q \ K	I	AM	QUAN
I	0.9	0.01	0.09
AM	0.14	0.81	0.05
QUAN	0.56	0.02	0.42

o Solution: Using more Attention Layers (Multi-Head)

→ The returned attention results will be more diverse and objective

Some examples of Attention in the image

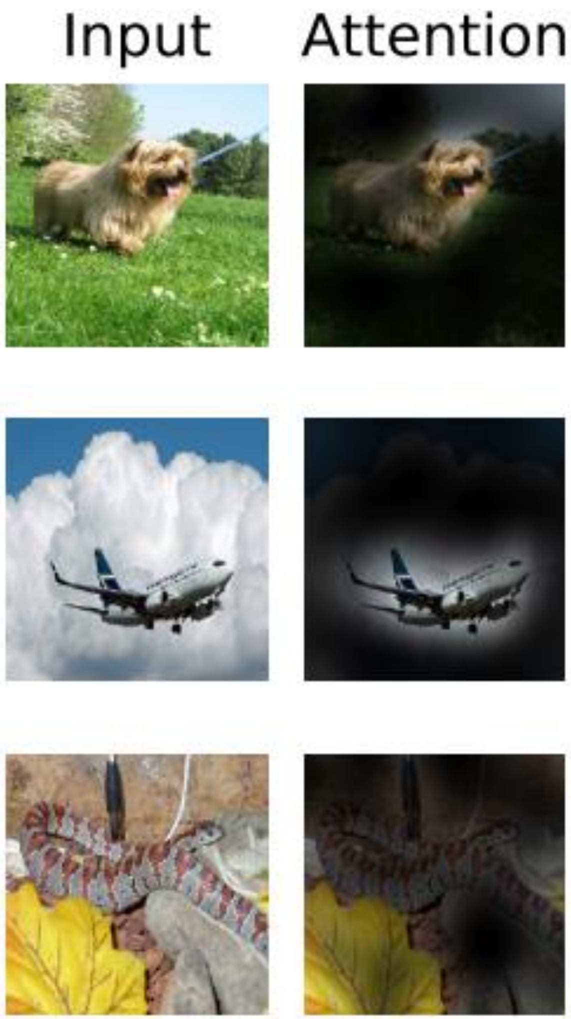


Figure 6. Representative examples of attention from the output token to the input space[4]

Results

- Evaluate the performance of the proposed MViT approach and compare it with concurrent and previous works on the same surface and invasive EEG databases.
- Evaluation based on performance metrics:
  - Accuracy (ACC)
  - Sensitivity (SENS - True Positive Rate)
  - Specificity (SPEC - True Negative Rate)
  - False-positive rate (FPR)[per hour]
  - Area under the ROC curve (AUC)

Results

1) MViT Prediction Performance on Surface Pediatric EEG

Achieve the best benchmark scores: SENS-99.8%(Highest), SPEC-99.7%(second-highest), ACC-99.8%(Highest), FPR-0.004/h(Lowest)

Authors	Year	EEG Features	Classifier	SENS (%)	SPEC (%)	ACC (%)	FPR (/h)
Zhang and Parhi [39]	2016	Spectral power	SVM	98.7	-	-	0.04
Cho et al. [40]	2016	Phase locking value	SVM	82.4	82.8	-	-
Usman et al. [41]	2017	Statistical and spectral moments	SVM	92.2	-	-	-
Khan et al. [23]	2018	Wavelet coefficients	CNN	86.6	-	-	0.147
Truong et al. [24]	2018	EEG Spectrogram	CNN	81.2	-	-	0.16
Tsiouris et al. [42]	2018	Spectral power, statistical moments	LSTM	99.3–99.8	99.3–99.9	-	0.02–0.11
Ozcan et al. [26]	2018	Spectral power, statistical moments	3D CNN	85.7	-	-	0.096
Zhang et al. [43]	2019	Common spatial patterns	CNN	92.0	-	90.0	0.12
Daoud et al. [44]	2019	Multi-channel time series	LSTM	99.7	99.6	99.7	0.004
Usman et al. [45]	2020	EEG Spectrogram + CNN features	SVM	92.7	90.8	-	-
Büyükçakır et al. [46]	2020	Statiscal moments, spectral power	MLP	89.8	-	-	0.081
Xu et al. [47]	2020	Raw EEG	CNN	98.8	-	-	0.074
Dissanayake et al. [48]	2021	Mel-frequency cepstral coefficients	Siamese NN	92.5	89.9	91.5	-
Hussein et al. [29]	2021	Scalogram	SDCN	98.9	-	-	-
Jana et al. [49]	2021	Raw EEG	CNN	92.0	86.4	-	0.136
Li et al. [50]	2021	Spectral-temporal features	GCN	95.5	-	-	0.109
Usman et al. [51]	2021	EEG Spectrogram	LSTM	93.0	92.5	-	-
Yang et al. [52]	2021	EEG Spectrogram	Residual network	89.3	93.0	92.1	-
Dissanayake et al. [53]	2022	Mel frequency cepstral coefficients	GNN	94.5	94.2	95.4	-
Gao et al. [54]	2022	Raw EEG	Dilated CNN	93.3	-	-	0.007
Zhang et al. [55]	2022	EEG Spectrogram	ViT	59.2–97.0	65.8–94.6	-	-
Proposed Method	2022	EEG Scalogram	MViT	99.8	99.7	99.8	0.004

Table 1. Benchmarking of the previous seizure-prediction methods and MViT approach: CHB–MIT EEG dataset[1]

Results

2) MViT Prediction Performance on Invasive Human and Canine EEG

Achieves the highest AUC score on the unseen data of the private test set*

(*) This paper's code was rerun by Kaggle on a private test set that is not provided to the author

Authors/ Team	Year	EEG Features	Classifier	SENS (%)	AUC Score Public/Private
Medrr [32]	2016	N/A	N/A	-	0.903/0.840
QMSDP [32]	2016	Correlation, Hurst exponent, fractal dimensions, Spectral entropy	LassoGLM, Bagged SVM, Random Forest	-	0.859/0.820
Birchwood [32]	2016	Covariance, spectral power	SVM	-	0.839/0.801
ESAI CEU-UCH [32]	2016	Spectral power, correlation, PCA	Neural Network, kNN	-	0.825/0.793
Michael Hills [32]	2016	Spectral power, correlation, spectral entropy, fractal dimensions	SVM	-	0.862/0.793
Truong et al. [24]	2018	EEG Spectrogram	CNN	75.0	-
Eberlein et al. [56]	2018	Multi-channel time series	CNN	-	0.843/-
Ma et al. [57]	2018	Spectral power, correlation	LSTM	-	0.894/-
Korshunova et al. [58]	2018	Spectral power	CNN	-	0.780/0.760
Liu et al. [27]	2019	PCA, spectral power	Multi-view CNN	-	0.837/0.842
Qi et al. [28]	2019	Spectral power, variance, correlation	Multi-scale CNN	-	0.829/0.774
Chen et al. [59]	2021	EEG Spectrogram	CNN	82.00	0.746/-
Hussein et al. [29]	2021	EEG Scalogram	SDCN	88.45	0.928/0.856
Usman et al. [60]	2021	statistical and spectral moments	Ensemble of SVM, CNN, and LSTM	94.20	-
Zhao et al. [61]	2022	Raw EEG	CNN	91.77–93.48	0.953–0.977/-
Proposed Method	2022	EEG Scalogram	MViT	90.28	0.940/0.885

Table 2. Benchmarking of the previous seizure-prediction methods and MViT approach: Kaggle/AES Seizure Prediction dataset[2]

Results

3) MViT Prediction Performance on Invasive Human EEG

Achieve superior seizure-prediction sensitivity of 91.15% and AUC score of 0.924 on the Public test set

Authors/ Team	Year	EEG Features	Classifier	SENS (%)	AUC Score Public/Private
Cook et al. [15] *	2013	Signal energy	Decision tree, kNN	33.67	-
Karoly et al. [62] *	2017	Signal energy, circadian profile	Logistic regression	52.67	-
Kiral-Kornek et al. [16] *	2018	EEG Spectrogram, circadian profile	CNN	77.36	-
Not-so-random -anymore [33]	2018	Hurst exponent, spectral power, distribution attributes, fractal dimensions, AR error, and cross-frequency coherence	Extreme gradient boosting, kNN, SVM	-	0.853/0.807
Arete Associates [33]	2018	Correlation, entropy, zero-crossings, distribution statistics, and spectral power	Extremely randomized trees	-	0.783/0.799
GarethJones [33]	2018	Distribution statistics, spectral power, signal RMS, correlation, and spectral edge	SVM tree ensemble	-	0.815/0.797
QingnanTang [33]	2018	Spectral power, spectral entropy correlation, and spectral edge power	Gradient boosting, SVM	-	0.854/0.791
Nullset [33]	2018	Hjorth parameters, spectral power, spectral edge, spectral entropy, Shannon entropy , and fractal dimensions	Random Forest, adaptive boosting, and gradient boosting	-	0.844/0.746
Reuben et al. [63]	2019	Preictal probabilities from the top 8 teams in [33]	MLP	-	0.815/-
Varnosfaderani et al. [64]	2021	Temporal features, statistical moments, and spectral power	LSTM	86.80	0.920/-
Hussein et al. [29]	2021	EEG Scalogram	SDCN	89.52	0.883/-
Zhao et al. [61]	2022	Raw EEG	CNN	85.19–86.27	0.914–0.933/-
Proposed Method	2022	EEG Scalogram	MViT	91.15	0.924/-

Table 3. Benchmarking of the previous seizure-prediction methods and MViT approach: Melbourne University AES/MathWorks/NIH Seizure Prediction dataset[3]

Discussion

- Through the results of the MViT method on the CHB-MIT surface EEG data set[1] :
  - Demonstrates the ability of this approach to provide robust seizure prediction performance on unseen EEG data recorded from new patients
- Through the results of the MViT method on 2 sets of invasive EEG data[2],[3]:
  - Proving that it can accommodate the variations in EEG data across different subjects or over time for the same subject

  → The MViT model is an excellent candidate for clinical and real-life settings

  - Relaxing the need for manually extracting domain-based features and also much faster in obtaining the results on unseen data

Limitations

Vision Transformer is more robust than convolutional and recurrent neural networks in terms of more generality and reliability, but it still has limitations

- Large-scale vision transformers can require intensive power and computational resources
- Limiting their deployment on resource-constrained devices such as brain-computer interface(BCI) and seizure warning systems
- Also quite challenging to interpret vision transformers' decisions (Eg. Visualizing the image regions with the greatest impact on the EEG classification performance)

Conclusion

- A multi-channel vision transformer (MViT) algorithm for the accurate prediction of epileptic seizures was proposed
  - ➢ The EEG signals were split into non-overlapping 10-second chunks.
  - ➢ The EEG chunks were converted into image-like representations called "scalograms" using continuous wavelet transform.
  - ➢ Using non-overlapping patches of fixed-size from the scalogram images to train the MViT algorithm.
  - ➢ MViT uses multiple branches to learn temporal-spectral features from various EEG channels at the same time.

→ Extensive experiments demonstrate that the proposed MViT model outperforms other neural network models for seizure prediction.

Reference

1. Shoeb, A.H. Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.
2. Brinkmann, B.H.; Wagenaar, J.; Abbot, D.; Adkins, P.; Bosshard, S.C.; Chen, M.; Tieng, Q.M.; He, J.; Muñoz-Almaraz, F.; Botella-Rocamora, P.; et al. Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain 2016, 139, 1713–1722.
3. Kuhlmann, L.; Karoly, P.; Freestone, D.R.; Brinkmann, B.H.; Temko, A.; Barachant, A.; Li, F.; Titericz, G., Jr.; Lang, B.W.; Lavery, D.; et al. Epilepsyecosystem.org: Crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. Brain 2018, 141, 2619–2630.
4. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv 2020, arXiv:2010.11929.
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.