Lightweight Transformer exhibits comparable performance to LLMs for Seizure Prediction: A case for light-weight models for EEG data

Paras Parani, Umair Mohammad and Fahad Saeed
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, FL 33172, USA
Emails: {ppara014, umohamma, fsaeed}@fiu.edu

***Abstract*—Predicting seizures ahead of time will have a significant positive clinical impact for people with epilepsy. Advances in machine learning/artificial intelligence (ML/AI) has provided us the tools needed to perform such predictive tasks. To date, advanced deep learning (DL) architectures such as the convolutional neural network (CNN) and long short-term memory (LSTM) have been used with mixed results. However, highly connected activity exhibited by epileptic seizures necessitates the design of more complex ML techniques which can better capture the complex interconnected neurological processes. Other challenges include the variability of EEG sensor data quality, different epilepsy and seizure profiles, lack of annotated datasets and absence of ML-ready benchmarks. In addition, successful models will need to perform inference in almost real-time using limited hardware compute-capacity. To address these challenges, we propose a lightweight architecture, called *ESPFormer,* whose novelty lies in the simple and smaller model-size and a lower computational load footprint needed to infer in real-time compared to other works in the literature. To quantify the performance of this lightweight model, we compared its performance with a custom-designed residual neural network (ResNet), a pre-trained vision transformer (ViT) and a pre-trained large-language model (LLM). We tested ESPFormer on MLSPred-Bench which is the largest patient-independent seizure prediction dataset comprising 12 benchmarks. Our results demonstrate that ESPFormer provides the best performance in terms of prediction accuracy for 4/12 benchmarks with an average improvement of 2.65% compared to the LLM, 3.35% compared to the ViT and 17.65% compared to the ResNet – and comparable results for other benchmarks. Our results indicate that lightweight transformer architecture may outperform resource-intensive LLM based models for real-time EEG-based seizure predictions.**

Keywords—Large Language Model (LLM), Vision Transformer (ViT), Electroencephalography (EEG), Epilepsy, Seizure Prediction

I. Introduction

Epilepsy is a chronic neurological disorder that affects over 65 million people globally [1] and causes repeat uncontrolled seizures. Seizure can be characterized with multiple symptoms including but not limited to simple loss of focus and short-term memory lost to loss of muscular control and unconsciousness resulting in falls and injuries [2]. At the fundamental level, a seizure is an increased activity in the brain that can potentially be tracked using electroencephalography (EEG). Design and development of machine-learning models that can predict a seizure before it happens using EEG data is an active area of research. Successful seizure prediction will give caregivers and people with epilepsy a tool that can alleviate significant financial and social costs. However, developing generalizable DL models is challenging due to multiple reasons including limited annotated data, variability due to different epilepsy and seizure types, varying EEG sensors and data-acquisition protocols, and co-morbidities.

Existing seizure prediction models are patient-specific and usually tested on the CHB-MIT dataset [3]. Recent works have used CNN-LSTM [4], 3D CNN-LSTM [5], and contrastive learning techniques [6] and achieve a prediction sensitivity up-to 95%. However, most of these patient-specific models do not utilize leave-one-out cross-validation (LOOCV) strategies which may over-estimate the models' reported performance due to data-leakage. This conjecture is recently confirmed by studies such as [7] where one of the most successful end-to-end seizure prediction systems based on CNN-LSTM [8], with a reported prediction sensitivity of 99.6%, is able to achieve only up-to 70% sensitivity when LOOCV is applied [7]. In addition, many of the models still report either low sensitivity or too many false positives which amplifies the need for new training and benchmarking approaches [9]. Whereas earlier DL models focused on CNN and LSTM pre-dominantly, the latest models leverage the Transformer architecture [10] [11]. Since a seizure may depend on interconnected neurological processes, which are recorded by the EEG data, newer models such as a large language model (LLM) [12], or a vision transformer (ViT) [13]–[15] may prove more adept at learning specific seizure patterns with the aid of attention mechanisms. To date, an end-to-end transformer configuration for seizure prediction has not been extensively studied. A few recent works [10][11] do apply the transformer architecture for feature extraction. However, these works [4]–[6], [8], [10], [11] [16], [17] are neither patient-independent nor tested on larger cohorts and testing is mostly done without LOOCV. This neither eliminates data leakage nor the associated problem of deceptively inflating performance. For effective assessment, datasets with disjoint training and testing sets are urgently needed as illustrated in our recent *MLSPred-Bench* benchmark [9]. Our previous work [18] investigated the use of pre-trained ViTs and LLMs for seizure

Paras Parani and Umair Mohammad have contributed equally to this work. Corresponding author can be contacted at fsaeed@fiu.edu.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

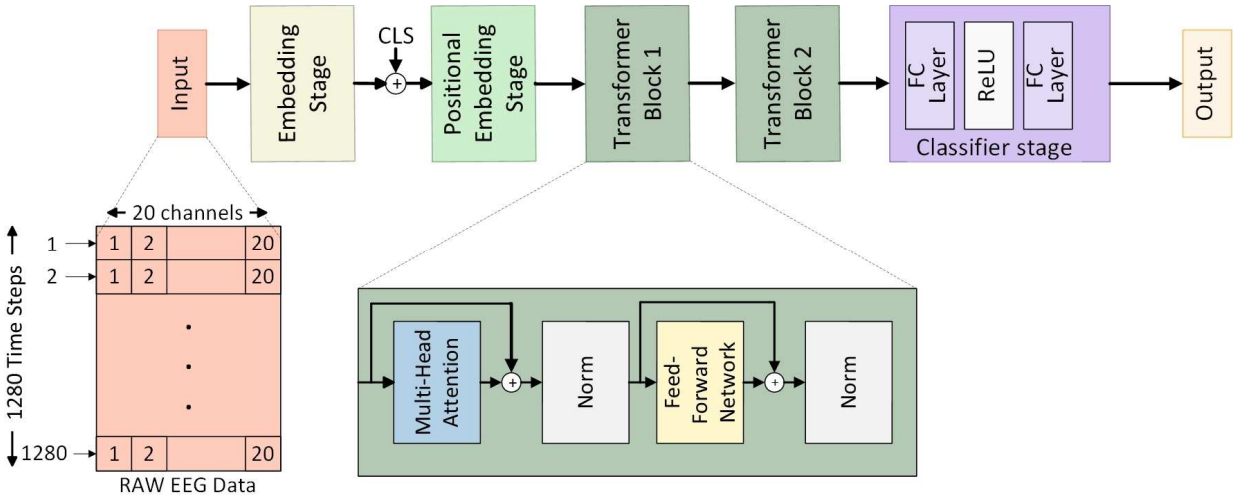

Fig. 1. Illustration of ESPFormer architecture. It contains an embedding stage, positional embedding and two transformer blocks.

prediction on a large cohort with reasonable performance advantages as compared to traditional models.

In this short paper, we design a simple transformer-based architecture to quantify the performance benefits and generalizability as compared to large LLMs and ViTs. To accomplish this, we solve two computational challenges: (1) creating an innovative technique for representing complex multi-channel EEG time-series data in a tokenized form; (2) developing a minimal transformer called ***E****pileptic* ***S****eizure* ***P****rediction* ***Former*** *(*hereafter: *ESPFormer)*. Solving these challenges then lend to the following solutions: (1) We develop wrangling methods that ***transform*** multi-dimensional time-series EEG data into a format suitable for ESPFormer. (2) We ***re-design*** the basic transformer architecture by modifying the embedding and positional embedding layers to accept and process multi-channel EEG data and by adding Fully Connected (FC) layers for classification. (3) We validate our strategy by comparing it to a pre-trained ViT and LLM on one of the ***largest patient-independent datasets*** for scalp EEG data [9]. Such comparison and quantification then can be useful for various scientists and researchers to select the kind of model that would be useful for their specific purpose.

II. Materials and Methods

A. Dataset and Tools

A seizure can be defined in four stages: ***preictal*** (before a seizure), ***ictal*** (main symptomatic phase), ***post-ictal*** (recovery stage) and ***interictal*** (periods of normal function between two seizures). Recall that the seizure detection task is simpler where the objective is to discriminate between ictal and non-ictal segments. In contrast, prediction is a much more challenging task which requires the discrimination between preictal and interictal segments. Since there is clinical uncertainty among the preictal stages of different subjects, usually a seizure prediction horizon (SPH) and a seizure occur period (SOP) are defined. The SPH corresponds to the preictal stage whereas the SOP is a small interval between the SPH and the start of the ictal phase. This gap time reflected by the SOP ensures prediction ahead of time. To deal with uncertainty of preictal phases, we recently defined 12 benchmarks called MLSPred-Bench [9] with the intention of promoting generalizable patient-independent models using a large seizure cohort. Based on the literature, the values of $SPH \in \{2, 5, 15, 30\}$ minutes and the $SOP \in \{1, 2, 5\}$ minutes were chosen for these benchmarks. The twelve corresponding benchmarks {BM1, BM2, …, BM12} are illustrated in Table 1 and more information can be obtained in our original paper [9]. All experiments reported in this paper were carried out on a single compute node with two CPUs and GPUs each. Each CPU has 10 cores whereas each GPU has 10,752 CUDA cores with up-to 96GB RAM. The Python programming framework and associated ML/DL libraries were used to develop and test all models.

B. ESPFormer Design

There are two major aspects to the novelty of ESPFormer: first, we develop an innovative strategy to prepare the EEG data for input to the transformer module and secondly, we re-design it to classify short EEG segments as preictal or interictal. This results in a simple, custom lightweight model capable of efficiently processing patient-independent EEG data and predicting epileptic seizures. The overall architecture is described in Figure 1. Each EEG data sample from all benchmarks in MLSPred-Bench comprises five-second duration segments. Each segment is either labeled as '1' representing the preictal class or '0' for the interictal class. The EEG data itself is collected from 20 channels sampled at 256 Hz. This is complex multi-channel time-series data with each sequence containing $5 \times 256 = 1280$ time-steps from 20 channels; resulting in sequence data of shape 20×1280. We conceptualize each data point as a single "sentence" containing 1280 "words", with each word embedded in a 20-dimensional space. However, rather than using tokens, we re-design the model so that it can accept raw EEG data sequences. This is implemented with the aid of an Embedding stage that maps the original sequence from 20-dimensional space to a 64-dimensional space, while maintaining the sequence length, using a fully connected layer as follows:

$$S' = W \times S + B \quad (1)$$

where $S' \in \mathbb{R}^{64\times1280}$ is the new sequence, $W \in \mathbb{R}^{64\times20}$ is the weight matrix, $S \in \mathbb{R}^{20\times1280}$ is the original sequence, and $b \in \mathbb{R}^{64\times1280}$ are the bias vectors. Then, we add a learnable token

TABLE I. BENCHMARK DESCRIPTIONSFROM MLSPRED-BENCH [9]

Benchmark (BM) Description		SOP		
		1 minute	*2 minutes*	*5 minutes*
SPH	*2 minutes*	BM1	BM2	BM3
	5 minutes	BM4	BM5	BM6
	15 minutes	BM7	BM8	BM9
	30 minutes	BM10	BM11	BM12

$cls \in \mathbb{R}^{1\times 64}$ by appending cls^T to the start of the sequence S' as shown in equation (2). SE serves as an aggregated representation of the sequence for the classification; $(.)^T$ is the transpose operation.

$$SE = [cls^T, S'] \tag{2}$$

The new sequence $SE \in \mathbb{R}^{64\times 1281}$ is passed through the Positional Embedding layer to learn the positional information of each word including the cls token, during training as follows:

$$Z = PE + SE \tag{3}$$

where $Z \in \mathbb{R}^{64\times 1281}$ is the final sequence with the positional information and $PE \in \mathbb{R}^{64\times 1281}$ represents the learnable positional embeddings. This positional embedding is essential and allows Transformers to learn from ordered sequences. The final sequence Z is then passed through a series of Transformer Encoder blocks. Each transformer block includes the multi-headed self-attention mechanism with four heads to capture detailed temporal relationships. The first step of the attention mechanism can be described as follows:

$$Q = ZW_Q; K = ZW_K; V = ZW_V \tag{4}$$

The input to the transformer block is Z, Q is the query, K is the key and V is the value, and W_Q, W_V, W_K are learnable projection matrices for Q, K and V respectively. These are used to calculate the attention score as follows:

$$X = softmax\left(\frac{QK^T}{\sqrt{D_K}}\right)V \tag{5}$$

Each attention block is followed by a normalization layer to obtain $X_{norm} = norm(Z + X)$. The normalized output is passed through two fully connected layers and a second normalization layer to get the encoded output Y_{out} as follows:

$$Y = W_2\big(GELU(W_1 X_{norm} + b_1)\big) + b_2 \tag{6}$$

$$Y_{out} = norm(X_{norm} + Y) \tag{7}$$

W_1 and W_2 are the weights of the first and second linear respectively, b_1 and b_2 are the biases. GELU is the gaussian error linear unit activation function. Each transformer block also includes a residual skip connection to stabilize learning. Finally, the classifier head consists of two linear layers with ReLU activation followed by a $softmax$ layer to generate the output probabilities. For training the classifier, only the cls token extracted from the last encoder output is used. We used following hyperparameters to train the model: a dropout probability of 0.1 in the transformer block, LR of 1e-4, warm-up ratio 0.1, weight decay of 1%, batch size of 32 and train for a total of 15 epochs.

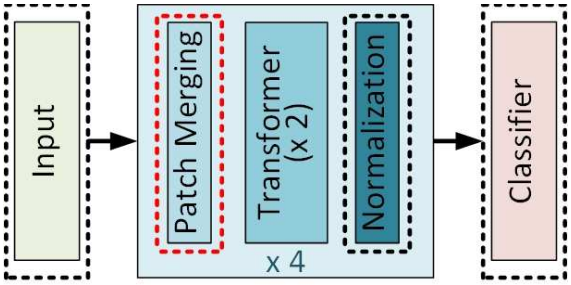

Fig. 2. Illustration of the six stages of the pre-trained ViT. There is an input layer, four successive transformer blocks followed by the classifier head. The re-trained layers are encapsulated in black dashed lines. Note that the patch merging layer encapsulated by the red dashed line is only re-trained in the first transformer stage.

C. ViT, LLM and ResNet designs

ViT Model. The ViT was adopted from the SegFormer [15] and comprises of six stages. The first stage is the input and the last stage is the classification phase. Each of the four stages in between contain a patch merging layer, two transformer blocks, and a normalization layer. The re-trained stages are illustrated in Fig. 2. The notable component is the transformer block which contains both an encoder and a decoder. Instead of positional embeddings, the encoder uses a combination of dense layers and 3×3 convolutions, where in the first stage, we obtain:

$$Z = CONV(FC(X_{in})) \tag{8}$$

where X_{in} is the feature from self-attention modules, FC represents the fully connected (FC) network, and $CONV$ is the convolution operation with a 3×3 kernel. The positional information is obtained by applying the second step as follows:

$$Y = FC(GELU(Z)) \tag{9}$$

where $GELU$ is the Gaussian Error Linear Unit activation function. For re-training the ViT, we use the NVIDIA/Mit-B0 checkpoint that only uses the encoder and is fine-tuned on ImageNet-1k. We use a batch size of 128, 5 epochs, LR set to $1e-3$, a warm-up ratio of 0.1 and 1% weight decay.

LLM Model. The LLM used is adopted from the Longformer [12] which uses two types of attention patterns as the underlying architecture in its transformer blocks. The first pattern uses a sliding window approach to calculate multiple stacked attention scores and in the second step, each pair of tokens is used to calculate a global attention for each sequence. These scores are then fed to a cls token for classification. For the LLM, we only re-train the classification phase with an LR of $5e-4$, warm-up ratio 0.03, weight decay of 1%, batch size 64 and train for 4 epochs. Because the LLM is restricted to 4096 tokens for each segment we classify each channel independently, and make a final decision only if the majority of the channels (> 10) predict a seizure for that segment.

Custom ResNet Model. The ResNet architecture used is based on SPERTL [19]. It contains an initial convolutional block followed by four residual blocks. Inside each residual block, there are two successive convolutional blocks. The first convolutional block inside each residual block is linked to the second convolutional block with a direct connection, and to the first convolutional block of the next residual block with a skip

TABLE II. COMPARISON OF ALL TECHNIQUES IN TERMS OF AUC SCORE, ACCURACY, SENSITVITY AND SPECIFICITY

BM	ResNet				ViT				LLM				ESPFormer			
	AUC	***Acc***	***Sen***	***Spe***	***AUC***	***Acc***	***Sen***	***Spe***	***AUC***	***Acc***	***Sen***	***Spe***	***AUC***	***Acc***	***Sen***	***Spe***
1	66.8%	61.8%	90.8%	32.9%	76.3%	69.7%	66.5%	72.8%	69.5%	71.5%	80.0%	63.2%	**79.8%**	74.3%	80.4%	68.2%
2	66.5%	50.0%	99.7%	0.4%	76.2%	70.7%	73.9%	67.5%	69.8%	70.3%	82.8%	57.6%	**77.6%**	72.8%	78.7%	67.0%
3	69.4%	55.1%	98.4%	11.7%	**76.1%**	68.6%	69.6%	67.8%	73.2%	74.3%	85.5%	63.1%	59.9%	62.4%	75.6%	49.1%
4	63.0%	57.8%	63.2%	52.4%	76.7%	67.5%	73.5%	61.5%	70.2%	71.2%	82.4%	59.8%	**82.7%**	75.5%	81.5%	69.5%
5	0.7%	57.4%	29.1%	85.7%	76.1%	68.3%	73.8%	62.7%	71.6%	70.9%	78.9%	62.5%	**80.4%**	74.0%	80.5%	67.5%
6	72.9%	60.0%	22.9%	97.1%	**77.0%**	70.3%	76.1%	64.5%	71.1%	71.3%	67.0%	75.5%	74.4%	68.6%	75.5%	61.6%
7	**81.1%**	53.0%	7.0%	99.0%	71.9%	67.0%	75.2%	58.8%	69.3%	67.6%	84.3%	50.7%	61.6%	60.5%	68.2%	52.7%
8	**83.6%**	51.0%	2.6%	99.3%	72.0%	65.3%	71.0%	59.4%	70.8%	70.5%	86.5%	54.4%	61.8%	60.8%	69.8%	51.7%
9	**79.2%**	49.9%	99.6%	0.2%	71.2%	64.7%	69.4%	59.9%	71.8%	73.5%	89.9%	57.5%	59.7%	57.8%	65.4%	50.3%
10	56.4%	50.0%	100.0%	0.0%	41.0%	46.4%	66.8%	26.0%	**62.1%**	66.6%	67.0%	66.2%	53.4%	53.3%	30.7%	75.8%
11	80.5%	50.0%	100.0%	0.0%	78.7%	71.2%	64.9%	77.5%	**87.6%**	78.2%	67.7%	88.7%	83.0%	76.6%	20.1%	95.4%
12	50.8%	50.0%	100.0%	0.0%	42.8%	50.1%	63.2%	37.1%	**60.4%**	64.1%	77.2%	50.8%	53.5%	54.0%	23.0%	84.9%
Avg.	64.2%	53.8%	67.8%	39.9%	69.7%	65.0%	70.3%	59.6%	**70.6%**	**70.8%**	**79.1%**	62.5%	69.0%	65.9%	62.5%	**66.1%**

connection. Each convolutional block contains a convolution layer, batch normalization, RELU activation and max pooling. The last residual block is followed by two FC layers for classification. In contrast to the pre-trained ViT and LLM which require significant manipulation of data, the ResNet requires only raw EEG data.

D. Evaluation Metrics

For evaluation, let us define the true positive (TP) as a correct seizure prediction, false positive (FP) as an incorrect seizure prediction, a true negative (TN) as correctly predicting no-seizure and a false negative (FN) as incorrectly predicting no-seizure. We evaluate all models by comparing the receiver operating characteristic (ROC) curves and the ROC area under the curve (AUC) scores. The AUC score allows us to compare the trade-off between the sensitivity and specificity offered by the model, where we define the sensitivity (Sen), specificity (Spe) and accuracy (Acc) in (10). Sensitivity measures the model's ability to correctly predict seizures, specificity measures its ability to identify the interictal phases whereas the accuracy is the ability of the model to correctly predict all labels.

$$Sen = \frac{T_p}{T_p+F_N}, Spe = \frac{T_N}{T_N+F_P} \text{ and } Acc = \frac{T_p+T_N}{T_p+T_N+F_P+F_N} \quad (10)$$

III. RESULTS

Table II compares AUC score, accuracy, sensitivity, and specificity for all the models including ResNet, ViT, LLM and the proposed ESPFormer for each benchmark. Overall, the best performance is provided by the LLM with an average AUC score of 70.6%, accuracy of 70.8%, 79.1% sensitivity and 62.5% specificity. The ESPFormer provides an average AUC score of 69.0%, an accuracy of 65.9%, sensitivity of 62.5% and a specificity of 66.1%. However, the proposed ESPFormer provides the best performance in the first four out of six benchmarks. These results demonstrate that a smaller transformer-based model such as ESPFormer has utility for seizure prediction compared to more resource-intensive models such as ViTs and LLMs. For example, ESPFormer provides AUC scores of 79.8%, 77.6%, 82.7% and 80.4% respectively for BM1, BM2, BM4 and BM5. In contrast, the best AUC score for BM3 and BM6 is provided by the VIT. The best AUC score for BM7-BM9 is provided by the ResNet whereas the LLM provides the best AUC scores for BM10-BM12.

In general, for the benchmarks for which ESPFormer does well, it provides a good accuracy with a reasonable trade-off between the sensitivity and specificity. For example, for BM5 ESPFormer has an 80.5% sensitivity and 67.5% specificity with an accuracy of 74.0%. In contrast, the LLM, ViT and ResNet each have a respective sensitivity of 78.9%, 73.8% and 29.1% and specificity of 59.8%, 61.5% and 85.7%. This is important because examining just one metric such as the specificity implies that the ResNet performs better, but because of its very low sensitivity it has negligible predictive power and provides a 57.4% accuracy compared to the accuracy of ESPFormer which is 74.0%. The accuracy of the LLM is 70.9% and the ViT is 68.4% by contrast. These trends also hold through for BM1, BM2 and BM4. Interestingly, compared to the ESPFormer in BM4, the LLM has a higher sensitivity of 82,4% versus 81.5% but a much lower specificity of 59.8% versus 69.5% leading to a lower accuracy of 71.2% compared to 75.5% of the ESPFormer. For other benchmarks where the ESPFormer does not perform well, the LLM provides the best performance in general for accuracy, sensitivity and specificity even where the best AUC score may be provided by other models such as the ViT in BM3 and BM6 and the ResNet in BM7-BM9. Overall, the results indicate that setting the prediction threshold as 0.5 influences the models to lean towards predicting a seizure (the positive class) which is reflected in higher sensitivity values compared to specificity. We can tune this by using *N*-fold CV and modifying the threshold based on metrics like the F-score.

IV. DISCUSSIONS AND CONCLUSION

The proposed *ESPFormer* has been able to outperform the ViT [18] in four out of the six benchmarks. Recall the structure of the benchmarks where BM1-BM3 have an SPH of two minutes and BM4 – BM6 use an SPH of 5 minutes. With a uniform segment size of 5 seconds each, the number of preictal and interictal segments are less compared to BM7-BM12, but are extracted from a larger number of seizures. Hence, the ViTs work well because they are able to train on a large number of short-term temporal features from which they can easily extract spatial features. ESPFormer is able to improve upon this

performance because it combines this benefit of ViTs with the ability of LLMs to extract temporal dependencies from longer sequences. The only exception to this are benchmarks BM3 and BM6. Recall that the SOP or the gap before a seizure is the highest at 5 minutes for these benchmarks compared to 1 minute for BM1 and BM4 and 2 minutes for BM2 and BM5. This indicates that ESPFormer is able to beat ViTs for smaller SOP's but such a small model can only extract preictal biomarkers closer to a seizure. The higher the SOP, the further away we move from the start of the seizures requiring earlier prediction which requires more sophisticated models. However, neuroscience informs us that such signatures may be closer to the seizures most of the time, for most of the subjects, making our proposed light-weight model very attractive for real-time predictive apps. BM7-BM9 approximately double the number of available preictal samples per seizures but with half the number of total seizures. This is problematic due to two issues for ViT where the dataset size is too small but the size of total temporal features per preictal sample becomes too large; the same applies to ESPFormer and the pre-trained LLM. However, the ResNet works best for these benchmarks due to its comparatively 'less' deep architecture and superior ability to work with longer sequences compared to a ViT or ESPFormer. Once we shift to BM10 – BM12 with SPH of 30 minutes, the number of preictal and interictal samples are sufficient for an LLM to provide the best AUC sore. However, in terms of a balanced performance among accuracy, sensitivity and specificity, LLM beats all other models for all benchmarks except BM1, BM2, BM4 and BM5 where our proposed ESPFormer is superior. This highlights the significance of our results and the great potential for smaller architectures based on simple transformers rather than very deep models with tens of layers and millions of trainable parameters. ESPFormer can also be combined with simple convolutional or residual layers to beat the LLM's performance on other benchmarks while restricting the trainable parameters in the thousands which is further motivation for more investigative studies. Lastly, the pre-trained LLM version required an average of 60 hours training time on 2 GPUs for simple partial re-training of the classifier head. In contrast, ESPFormer requires only a few minutes for tens of epochs using similar resources.

Our proposed architecture based on a simple transformer provided the best performance on one-third of the benchmarks and came second only to the LLM on one of the largest patient-independent seizure prediction dataset (MLSPred-Bench). Further, it requires considerably less tuning resources compared to large pre-trained models with billions of parameters. In conclusion, our results show that smaller transformer-based architectures have significant potential for developing robust generalizable epileptic seizure prediction models.

ACKNOWLEDGMENT

This material is based upon work supported by the NSF grant TI-2213951 and OAC-2312599. Fahad Saeed was in addition supported by NIH R35GM153434. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or the NIH.

REFERENCES

[1] World Health Organization, 'Epilepsy', *Fact Sheets | Details*. 2023 [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy [Accessed: 17 June 2024].

[2] N. Mühlenfeld et al., 'Seizure related injuries – Frequent injury patterns, hospitalization and therapeutic aspects', *Chin. J. Traumatol. - Engl. Ed.*, vol. 25, no. 5, pp. 272–276, Sep. 2022 [Online]. Available: 10.1016/j.cjtee.2021.10.003.

[3] A. H. Shoeb and J. Guttag, 'Application of Machine Learning To Epileptic Seizure Detection', in *ICML*, 2010, pp. 975–982 [Online]. Available: https://icml.cc/Conferences/2010/papers/493.pdf.

[4] Y. Ma et al., 'A Multi-Channel Feature Fusion CNN-Bi-LSTM Epilepsy EEG Classification and Prediction Model Based on Attention Mechanism', *IEEE Access*, vol. 11, pp. 62855–62864, 2023 [Online]. Available: 10.1109/ACCESS.2023.3287927.

[5] X. Lu et al., 'An Epileptic Seizure Prediction Method Based on CBAM-3D CNN-LSTM Model', *IEEE J. Transl. Eng. Health Med.*, vol. 11, pp. 417–423, 2023 [Online]. Available: 10.1109/JTEHM.2023.3290036.

[6] L. Guo et al., 'CLEP: Contrastive Learning for Epileptic Seizure Prediction Using a Spatio-Temporal-Spectral Network', *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3915–3926, 2023 [Online]. Available: 10.1109/TNSRE.2023.3322275.

[7] Z. Wang et al., 'Power efficient refined seizure prediction algorithm based on an enhanced benchmarking', *Sci. Rep.*, vol. 11, no. 23498, Dec. 2021 [Online]. Available: https://doi.org/10.1038/s41598-021-02798-8.

[8] H. Daoud and M. A. Bayoumi, 'Efficient Epileptic Seizure Prediction Based on Deep Learning', *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 804–813, Oct. 2019 [Online]. Available: 10.1109/TBCAS.2019.2929053.

[9] U. Mohammad and F. Saeed, 'MLSPred-Bench: ML-Ready Benchmark Leveraging Seizure Detection EEG data for Predictive Models', *BioRxiv Prepr. BioRxiv202407176040006*, pp. 1–13, 2024 [Online]. Available: 10.1101/2024.07.17.604006.

[10] L. Xia et al., 'Hybrid LSTM-Transformer model for the prediction of epileptic seizure using scalp EEG', *IEEE Sens. J.*, vol. 24, no. 13, pp. 21123–21131, Jul. 2024 [Online]. Available: 10.1109/JSEN.2024.3401771.

[11] S. Shi and W. Liu, 'B2-ViT Net: Broad Vision Transformer Network with Broad Attention for Seizure Prediction', *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 178–188, 2024 [Online]. Available: 10.1109/TNSRE.2023.3346955.

[12] I. Beltagy et al., 'Longformer: The Long-Document Transformer', *ArXiv Prepr. ArXiv200405150*, pp. 1–17, Apr. 2020.

[13] A. Dosovitskiy et al., 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', in *International Conference on Learning Representations*, 2021, pp. 1–21 [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[14] Z. Liu et al., 'Swin Transformer V2: Scaling Up Capacity and Resolution', *Proc. 2022 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 11999–12009, 2022 [Online]. Available: 10.1109/CVPR52688.2022.01170.

[15] E. Xie et al., 'SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers', in *2021 Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021, pp. 1–14 [Online]. Available: https://openreview.net/forum?id=OG18MI5TRL.

[16] T. Dissanayake et al., 'Deep Learning for Patient-Independent Epileptic Seizure Prediction Using Scalp EEG Signals', *IEEE Sens. J.*, vol. 21, no. 7, pp. 9377–9388, Apr. 2021 [Online]. Available: 10.1109/JSEN.2021.3057076.

[17] T. Dissanayake et al., 'Geometric Deep Learning for Subject-Independent Epileptic Seizure Prediction using Scalp EEG Signals', *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 527–538, 2021 [Online]. Available: 10.1109/JBHI.2021.3100297.

[18] P. Parani et al., 'Utilizing Pretrained Vision Transformers and Large Language Models for Epileptic Seizure Prediction', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2024.11.03.621742, Nov. 2024 [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2024.11.03.621742.

[19] U. Mohammad and F. Saeed, 'SPERTL: Epileptic Seizure Prediction using EEG with ResNets and Transfer Learning', in *BHI-BSN 2022 - IEEE-EMBS International Conference on Biomedical and Health Informatics s, Symposium Proceedings*, 2022, pp. 1–5 [Online]. Available: https://ieeexplore.ieee.org/document/9926767.