scientific reports

Check for updates

OPEN

Exploration of an intrinsically explainable self-attention based model for prototype generation on single-channel EEG sleep stage classification

Brenton Adey, Ahsan Habib ✉ & Chandan Karmakar

Prototype-based methods in deep learning offer interpretable explanations for decisions by comparing inputs to typical representatives in the data. This study explores the adaptation of SESM, a self-attention-based prototype method successful in electrocardiogram (ECG) tasks, for electroencephalogram (EEG) signals. The architecture is evaluated on sleep stage classification, exploring its efficacy in predicting stages with single-channel EEG. The model achieves comparable test accuracy compared to EEGNet, a state-of-the-art black-box architecture for EEG classification. The generated prototypical components are exaimed qualitatively and using the area over the perterbation curve (AOPC) indicate some alignment with expected bio-markers for different sleep stages such as alpha spindles and slow waves in non-REM sleep, but the results are severely limited by the model's ability to only extract and present information in the time-domain. Ablation studies are used to explore the impact of kernel size, number of heads, and diversity threshold on model performance and explainability. This study represents the first application of a self-attention based prototype method to EEG data and provides a step forward in explainable AI for EEG data analysis.

Electroencephalogram (EEG) is a non-invasive method for recording brain activity, capturing electrical signals through electrodes placed on the skull and is a popular diagnostic tool for neuro-psychological disorders[1]. Extracting useful features from EEG signals is particularly difficult because it is often required to analyse them from time, frequency, and spatial domains[2]. In addition, EEG signals contain a lot of noise causing interference. This complexity makes the analysis of EEG more involved than other bio-signal data such as electrocardiogram (ECG).

Deep learning models deployed in sensitive domains such as healthcare require transparency in their decision-making so that clinicians can reason the predictions generated by these models[3,4]. The field of explainable artificial intelligence (XAI) attempts to provide techniques for facilitating this transparency[5]. An explanation in XAI is a comprehensible interface between a decision maker (e.g. model) and a human[6]. There is a limited XAI research for EEG analysis incorporating explainability techniques of some kind[7].

There is a distinction between post-hoc explainability, methods used to generate explanations for a model after it has been trained, and intrinsic methods of explainability; that is, creating *white-box* models that offer direct human-interpretable explanations for their predictions. The most common methods used on deep learning models for EEG involve saliency maps. These are visual representations that highlight the most important or relevant features in given input data, offering insights into the contribution of individual elements to a model's output[7,8]. Specific methods or variations such as layer-wise relevance propagation (LRP)[9] and class-activation maps (CAM)[10] have also been successfully applied to EEG[11,12]. It has been shown that DeepLift[13] best incorporates the temporal, spectral, and spatial domains compared to other visualisation-based methods, and has been used to provide explainations for state-of-the-art models such as EEGNet[7,14].

Perturbation-based approaches are another post-hoc technique used to generate neural network explanations which involve making small changes or perturbations to the input data and observing how these changes affect the model's predictions[15]. Some applications of these methods are able to make controlled perturbations of the both spectral and spatial features for a more useful application to EEG, though the explanations generated are global and not tied to a specific input[16,17].

School of Information Technology, Deakin University, Geelong 3225, Australia. ✉email: habib@deakin.edu.au

Although the post-hoc methods above provide a level of interpretability and transparency to deep learning architectures, these methods often come with inherent drawbacks. Notably, saliency, which is among the most common method for EEG deep learning models, suffers from a lack of model and input sensitivity for explanations in the temporal, spectral, and frequency domains[7,17]. Likewise, perturbation methods can generate out-of-distribution samples that are not reflective of what a classifier has truly learned[18]. Instead, there has been an increased focus on creating intrinsically explainable or *white-box* models[18]. These methods use the same mechanism for predictions and explanation generation, thus making the decision process transparent to even non-data experts.

Of these intrinsically explainable methods, constrained feature extraction has shown successful applications to electroencephalogram data. These methods replace the initial convolutional layers with ones that have fewer parameters and possess an interpretable meaning, such as Morlet wavelet-based kernel[2] and sinc-convolution layers[19–21]. However, by design, these methods are restricted in the features they can learn from the data[2,22].

The second major group of intrinsically explainable models are concept-based models. Concept-based models attempt to learn semantically distinct concepts from data which are then used for prediction generation and explanation. Prototype-based methods are a subclass of this group that build semantically unique concepts by matching them to concrete samples in the data[23]. Various approaches have been introduced to apply prototype-based methods to time-series data such as ProSeNet and SCN_{PRO}[22,24,25]. The self-explaining selective model (SESM), based on the self-explaining neural network (SENN), used a multi-head self-attention mechanism to select prototypical parts of sequences that were not necessarily continuous[26,27]. In contrast to other prototype-based time series approaches, the prototypes were sampled from the original data and did not require projection back onto the original space which increased interpretability[28,29]. The SESM model demonstrated a high level of accuracy for electrocardiogram tasks, outperforming ProSeNet and black-box architectures. Importantly, the interpretability scores for SESM were higher than other approaches. This success in ECG tasks highlights it's potential for application to EEG analysis.

Although, due to the longer segment length and more varied features of EEG compared to ECG, this success in ECG may not necessarily translate to EEG. The current work aims to assess the validity of the SESM architecture on EEG data for a sleep stage classification, assessing the impact on classification accuracy and analysing the spectral and temporal features of the generated prototypes as a preliminary assessment of their interpretability. To the authors' knowledge, this work represents the first application of a self-attention prototype-based method to EEG data and contributes to discussions of its viability in the domain.

Methodology

Problem formulation

Sleep stage classification using EEG aims to automatically categorise distinct stages of sleep based on electrical impulses detected by electrodes located on the surface of the scalp. Accurate classification of sleep stages, including wakefulness, NREM (non-rapid eye movement) sleep, and REM (rapid eye movement) sleep, is fundamental for diagnosing sleep disorders and comprehending the intricate patterns of sleep. Sleep state classification is a well researched application of EEG analysis which makes it a useful baseline for assessing the viability of SESM. The challenges in sleep stage classification using EEG include inter-subject variability and noise in EEG recordings, necessitating the development of robust and generalisable feature extraction.

Various biomarkers across the temporal and frequency domains play a pivotal role in sleep stage classification. These include characteristics such as sleep spindles, K-complexes, slow-wave activity, power spectral density, and frequency bands (e.g., delta, theta, alpha, beta, and gamma)[30–33]. Figure 1 visualises typical examples of these wave patterns found during the various stages of normal sleep[34], Figure 10.2. Sleep stage classification using EEG typically involves the utilisation of multiple EEG channels to capture a comprehensive view of brain activity, also known as spatial features[30]. However, there has been success in performing sleep stage analysis while only considering single-channel EEG from the Fpz-Cz or Pz-Oz electrodes[35].

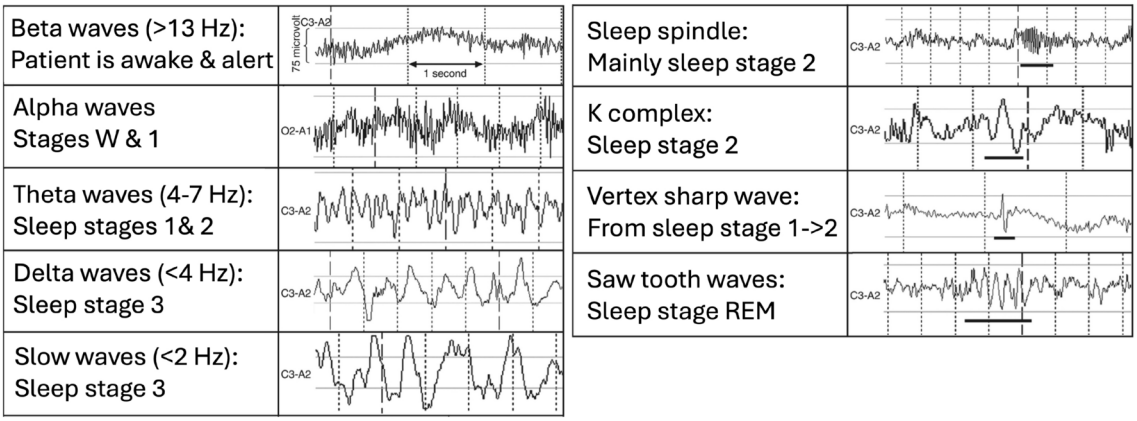


Figure 1. Example EEG waveforms found in various stages of normal sleep. Adapted from[34], Figure 10.2.

For the current research, these bio-markers are considered the target concepts. The goal is to encode the raw EEG signal from a single channel and produce an output that automatically identifies sub-sections of the signal with these concepts present. These concepts are used as prototypes for the downstream sleep stage classification task. The final model represents a method that can learn intrinsically explainable concepts from the data and could be generalised to inter-subject and intra-subject scenarios or applied to different EEG-related classification tasks. The current work considers single-channel EEG data as it is the most analogous to single-channel ECG data discussed in previous works. However, as discussed, a single-channel EEG presents a greater challenge for analysis, even with spatial information disregarded, due to the greater signal-to-noise ratio and variety of features across the time and frequency domains.

Network architecture

The proposed solution adapts the SESM architecture for use on single-channel EEG data[26]. Figure 2 shows the architecture of this solution.

First, the single-channel EEG time series is embedded into a latent d_{embed}-dimensional space through the embedder ($\mathscr{E}$) a 1D convolutional layer. The components of the embedded representation X are used as the smallest sequence units for the following layers.

As with the SESM architecture, there are three major components. The conceptizer $\mathscr{C}$ first contains the multi-headed self-attention mechanism which generates a binary vector s_h of selective actions from the embedded EEG signal for every h^{th} head for all H number of heads. The vector s_h is the same length as the sequence length N. The equation for the attention mechanism is $s_h = \text{Gumbel-Sigmoid}\left(\frac{QK^TW}{\sqrt{d_h}}\right)$.

Where $Q, K \in \mathbb{R}^{N\text{x}H}$ are matrices queries and keys, and $W \in \mathbb{R}^{N\text{x}1}$ is a matrix to ensure pair-wise attentions in s_h are longer element-wise attentions once binarised. Following this, the selective actions s_h are applied to the embedded signal X and encoded into further hidden representations through three sequential blocks of 1-D CNN layers with batch normalisation and 1-D max pooling. The result is H matrices $c_h \in \mathbb{R}^{B\text{x}m}$ representing concepts where B is the batch size, or number of samples, and m is the arbitrary number of hidden representations. In the proposed design, because the embedded signal aggregates all channels, concepts selected by the heads also represent interactions between all channels. Hence, it is expected that a greater number of heads will be required to fully capture the semantic information.

The paramaterizer $\mathscr{P}$ consists of three sequential blocks of 1-D CNN layers with batch normalisation and 1-D max pooling followed by a fully connected layer activated by a soft-max function which projects the CNN representations to the number of prototypical parts. In essence, this network acts as a task-agnostic model for predicting "relevance weights" for each prototypical part. The result is a H-dimensional vector P of scalars where each $p_h \in P$ is the "relevance" weight for the h^{th} head.

Finally, the aggregator $\mathscr{G}$ combines the outputs of $\mathscr{C}$ and $\mathscr{P}$ by first applying a fully connected layer to the output of $\mathscr{C}$ where the number of output nodes is the number of classes in the upstream task. Next the "relevance weights" from $\mathscr{P}$ are applied to the output through matrix multiplication.

In summary, $\mathscr{E}$ embeds a single-channel EEG signal into a latent representation X. $\mathscr{C}(X)$ outputs H concept matrices for X, and $\mathscr{P}(X)$ outputs "relevance weights" X. $\mathscr{G}$ combines these scores and outputs class-wise activations for a given classification task. This process is summarised in Equation 1 where x is a single-channel EEG signal and H represents the number of heads, or equivalently the number of concepts.

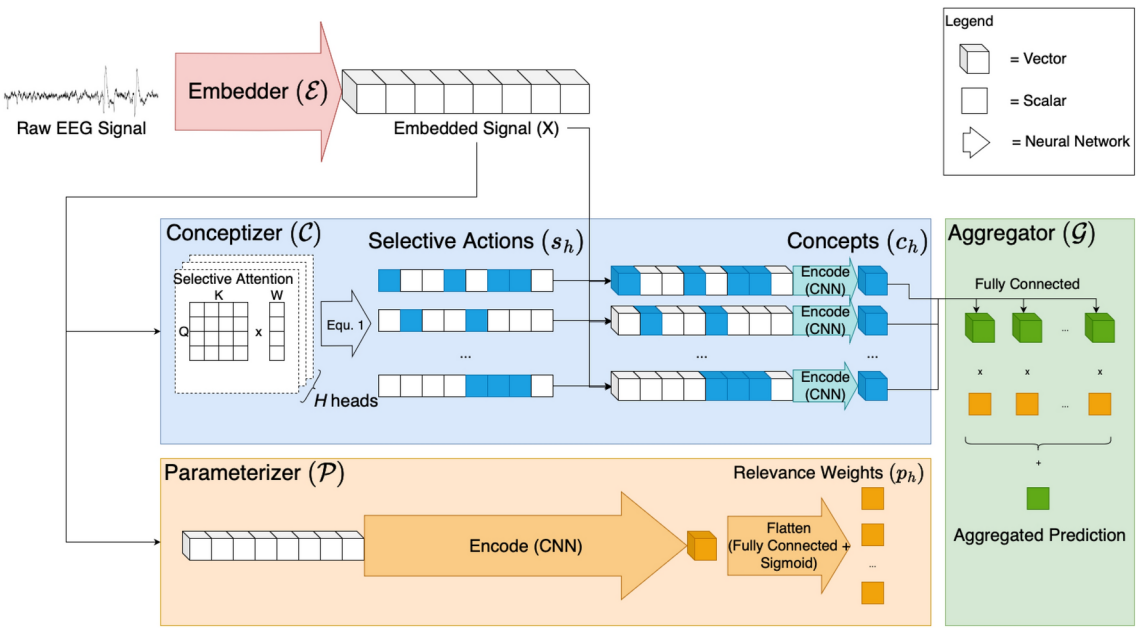


Figure 2. Proposed model architecture for self-selecting prototypical-parts model for EEG based off[26].

$$\text{EEG-SESM}(x) = \mathscr{G}\Big(\mathscr{P}\big(\mathscr{E}(x)\big)\mathscr{C}\big(\mathscr{E}(x)\big)\Big) = \mathscr{G}\big(\mathscr{P}(X)\mathscr{C}(X)\big) = \sum_{h=1}^{H} p_h c_h \tag{1}$$

The learning criteria aim to balance the accuracy of the predictions and the interpretability of the prototypes. Unique to SESM however is the way the cost-function is coupled to the outputs of the three major architectural components: The cross-entropy loss is used to capture classification accuracy, a diversity term is leveraged to ensure prototypical parts are sufficiently different and select different parts of the input sequence, a stability regularisation term guides each head of the conceptizer to capture one and only one concept by minimising the cosine distance between the encoded concepts from the same head, and an additional locality term prevents the attention heads from selecting long sequences in the conceptizer.

The exact formulations of the diversity ($\mathscr{L}_d$), stability ($\mathscr{L}_s$), and locality ($\mathscr{L}_l$) terms are given in equations below.

$$\begin{aligned}
\mathscr{L}_d &= \sum_{i=1}^{H-1}\sum_{j=i+1}^{H}\Big[\text{RELU}(d_{min} - \|s_i - s_j\|^2)\Big] \\
\mathscr{L}_s &= \sum_{h=1}^{H}\sum_{i=1}^{B-1}\sum_{j=i+1}^{B}\Big[1 - d_{cos}(c_i^h, c_j^h)\Big] \\
\mathscr{L}_l &= \sum_{h=1}^{H}\frac{1}{N}\sum_{i=1}^{N} s_{h,i}
\end{aligned} \tag{2}$$

where d_{min} in $\mathscr{L}_d$ is a threshold value, B in $\mathscr{L}_s$ is the batch size and $d_{cos}(c_i^h, c_j^h)$ is the cosine similarity. Hence the total loss is defined as in equation 3, where λ_d, λ_s, and λ_l are hyperparameter weight values for diversity, stability, and locality respectively.

$$\mathscr{L} = \mathscr{L}_{\text{cross-entropy}} + \lambda_d\mathscr{L}_d + \lambda_s\mathscr{L}_s + \lambda_l\mathscr{L}_l \tag{3}$$

Of note however is that there must be a two-stage training process, one to train $\mathscr{E}$, and another where the weights of the $\mathscr{E}$ are fixed and the remaining model is trained. This is to ensure the embedded representations remain consistent through training while the prototypical concepts are being learnt.

Table 1 shows the layer details for the entire model, including output shape and number of parameters. Note that the input mask to remove zero-valued inputs present in the original SESM implementation for ECG has not been included as it is not relevant for EEG data.

Dataset

The dataset used in the current research is the Sleep Cassette Data from the Sleep EDF Database[36] sourced from PhysioNet[37]. The data is available through the following link https://www.physionet.org/content/sleep-edf/1.0.0/. The data consists of overnight polysomnographic recordings obtained from cassette-based sleep studies. In the current research, only the EEG data from the Fpz-Cz channel is used.

Note that although an expanded version of the dataset was introduced, the current research used the original version for simplicity and ease of iteration. The data comprises of 39 recordings from 20 healthy participants during 24 hours in their normal daily life sampled at 100 Hz. Due to the high number of awake hours in the original recordings, only wake periods of 30 minutes before and after the sleep periods were considered.

The dataset was organised into 30-second epochs, maintaining the 100 Hz resolution as in similar studies[38]. Additionally, data from underrepresented classes was over-sampled to reduce class bias during training. In all other regards, the signal was unmodified from its raw state. The data labels consist of 5 classes: Awake (W), non-REM sleep stage 1 (N1), non-REM sleep stage 2 (N2), non-REM sleep stage 3 (N3), and REM (rapid eye movement) sleep.

Experiments

EEG-SESM was implemented in Python using PyTorch as the framework. The code for the implementation can be found at https://github.com/BrentonAD/EEG-SESM. EEG-SESM was trained on a virtual machine running Windows 10 with an Nvidia GeForce GTX 1080Ti 11GB GPU, an Intel Xeon CPU, and 128GB of shared random access memory. Additionally, a Tensorflow implementation of EEGNet[14] was used for a model comparison. This model was trained on the same dataset on an Apple Macbook Pro with an M1 Pro SoC with 32GB of RAM and Metal GPU acceleration.

In all experiments, d_{embed} and d_{hidden} were fixed to 64, the dropout before the final layer of $\mathscr{P}$ and optimiser initial learning rate were set at 0.2 and $1e{-}3$ respectively. The Adam optimiser was used in the training of the models with no weight decay, $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and a scheduled learning rate decrease of 95% per epoch after the first five epochs[39]. Training was terminated according to an early stopping policy which was executed when the total loss 3 remained consistent for 15 consecutive epochs.

To accommodate the longer sequence length, lower sampling rate, and other differences in EEG signals hyperparameters were optimised before model evaluation. The optimal hyperparameters were derived through a non-exhaustive search maximising accuracy on the validation dataset. Some of these optimal hyperparameters include - number of heads (H): 5, Embedded dimension (d_{embed}): 64, convolution Kernel size (k): 40, Minimum

Component	Layer type	Output shape	Kernel shape	Number of parameters
Embedder	Conv1d	[16, 64, 3000]	[40]	2,624
	BatchNorm1d	[16, 64, 3000]	–	128
	Swish	[16, 64, 3000]	–	–
Conceptizer	Linear	[16, 3000, 60]	–	3,900
	Linear	[16, 3000, 60]	–	3,900
	Linear	[16, 3000, 5]	–	325
	GumbelSigmoid	[16, 5, 3000, 1]	–	–
	ConvNormPool	[80, 64, 1500]	[40]	–
	Conv1d	[80, 64, 2961]	[40]	163,904
	BatchNorm1d	[80, 64, 2961]	–	128
Parameterizer	Conv1d	[16, 64, 2961]	[40]	163,904
	BatchNorm1d	[16, 64, 2961]	–	128
	Swish	[16, 64, 2961]	–	–
	Conv1d	[16, 64, 2961]	[40]	163,904
	BatchNorm1d	[16, 64, 2961]	–	128
	Swish	[16, 64, 2961]	–	–
	Conv1d	[16, 64, 2961]	[40]	163,904
	BatchNorm1d	[16, 64, 2961]	–	128
	Swish	[16, 64, 2961]	–	–
Aggregator	BatchNorm1d	[80, 64]	–	128
	Swish	[80, 64]	–	–
	Linear	[80, 5]	–	325
Total params: 1,980,167				
Trainable params: 1,980,167				
Non-trainable params: 0				

Table 1. Layer details of the model.

threshold (d_{min}) for diversity: 9, weight factor for diversity (λ_d)/stability (λ_s)/locality (λ_s): 1.0/0.2/0.2, and initial learning rate: $1e{-}3$.

To provide a reasonable baseline, the experiments are compared with EEGNet[14], a state-of-the-art deep learning model for EEG data analysis. Numerous iterations of the EEGNet training were run to tune the hyperparameters. The best results were found using a dropout rate of 0.5, kernel Length of 64, and 64 temporal and pointwise filters. The EEGNet model was trained with an early termination policy monitoring the validation loss, with a patience of 20 and starting from epoch 50.

The primary goal of the current work is not to assess the cross-subject generalisability of the model performance however to minimise the influence of any inherent bias in the data, the model was trained and tested on three random splits of the data, each with 15% of the subjects held out for testing, 10% of the total for validation and hyperparameter selection, and remaining data used for model training. The final results are given both individually for each test split and averaged across all test splits. Importantly, no data used to train the model or determine it's hyperparameters was used to evaluate the models performance. The implementation of EEGNet was trained and evaluated on the same data. Visualisation of the prototypical parts was performed using a randomly selected test subject from one of the splits mentioned above.

In addition, three ablation studies were performed to evaluate the impact of varying the kernel size, number of heads, and diversity threshold on performance and explainability. In these studies only split 1 was used. For the kernel size and number of heads studies, a diversity threshold (d_{min}) of 2 was used. For the diversity threshold study the number of heads was set to 4.

Effects of kernel size, number of heads and diversity thresholds were observed following below steps:

1. To assess the influence of kernel size (k) on the predictive power and explainability the model was trained with kernel sizes of 10, 20, 30, 40, 50, and 60,
2. The number of heads (H) was varied from 4 to 9 to assess the impact on the accuracy and explainability of the model, and
3. The diversity threshold d_{min} controls the minimum L^2 distance two selective actions (such as, the output of each head) are permitted to be. A distance greater than d_{min} will result in a diversity loss of 0. The effect of d_{min} on the performance of the model was assessed for the values 2, 5, 7, 9, 10, and 11.Similar to previous studies, the area over perturbation curve (AOPC) was used for objective evaluation of the model explainability[26]. In the current work, this is defined in equation 4

Model	Accuracy	Precision	Recall
EEG-SESM	**0.764** ± 0.014	**0.701** ± 0.013	0.693 ± 0.019
EEGNet	0.716 ± 0.007	0.686 ± 0.023	**0.727** ± 0.019

Table 2. Mean and standard deviation of accuracy, precision, and recall across all three hold-out test dataset for EEG-SESM and EEGNet. Significant values are in bold.

Model	W	N1	N2	N3	REM
EEG-SESM	**0.738** ± 0.092	0.286 ± 0.019)	**0.851** ± 0.027	0.863 ± 0.111	**0.724** ±0.091
EEGNet	0.670 ± 0.138	**0.570** ± 0.045	0.711 ± 0.088	**0.990** ± 0.008	0.695 ± 0.111

Table 3. Mean and standard deviation of class-wise accuracies for EEG-SESM and EEGNet across all three data splits. Significant values are in bold.

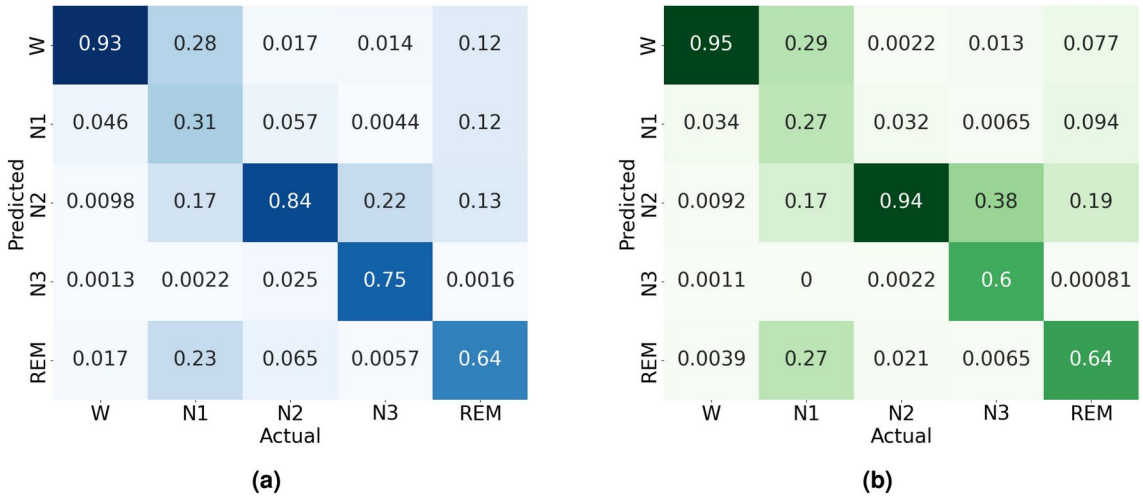


Figure 3. Confusion matrices on test subjects for all three data splits for a) EEG-SESM and b) EEGNet models.

$$\mathrm{AOPC} = \frac{1}{H-1}\left\langle \sum_{h=1}^{H-1} f(x) - f(x_{\setminus 1,\ldots,h}) \right\rangle \tag{4}$$

where H is the number of heads, $f(x)$ is the probability of predicting the true label, $f(x_{\setminus 1,\ldots,h})$ is the probability of predicting the true label excluding the top h heads sorted by relevance, and $\langle\cdot\rangle$ is the average over all the samples. The maximum value of AOPC is 1, a higher AOPC suggests that the model changes its prediction more significantly when heads are removed, indicating that the prototypes contributing most to the final prediction are more significant and have less redundancy.

Results

Model performance

The mean and standard deviation test accuracy, precision, and recall across all three splits for both models are shown in Table 2. Overall EEG-SESM demonstrated a comparable standard deviation of metrics across the test data splits. The reported accuracy and precision for EEG-SESM were much greater than EEGNet, though EEGNet reported a slightly higher recall.

The mean class-wise accuracies across all three splits are given in Table 3. Both classes reported the highest accuracy for the N3 class, however EEG-SESM showed a much higher variance for this class. EEG-SESM seemingly outperformed EEGNet for W, N2 and REM, displaying much higher averages with lower standard deviations. Conversely, EEGNet was stronger at predicting N1 and N3. The average accuracy for the N1 class with EEG-SESM appears notably low, even compared to the results from EEGNet. Figure 3a shows the confusion matrix on the test data for EEG-SESM, which suggests that a majority of the N1 samples were misclassified as W or REM. Other classes do not exhibit a miss-classification bias as strong as this. The W class showed the highest proportion of false positives. The confusion matrix for EEGNet, shown in Fig. 3b demonstrates a similar pattern of N1 miss-classifying N1 as W or REM.

To help interpret the results in a more rigorous way, the distributions shown in Tables 2 and 3 were compared using a two-tailed T-test between EEG-SESM and EEGNet. The p-values obtained from these comparisons were then corrected for multiple tests using the Benjamini-Hochberg false discovery rate procedure[40]. The corrected two-tailed t-test p-values accuracy, precision, and recall were 0.019, 0.365, 0.138 respectively. Similarly, the corrected values for the W, N1, N2, N3, and REM class-wise accuracies were 0.646, 0.003, 0.147, 0.203, 0.737.

Hence, any apparent differences in precision, recall, and most class-wise accuracies are not evidence that the evaluation results from the two models are statistically sampled from different distributions (to a significance factor of 5%). The only statistically significant results were that of the mean accuracy and N2 class-wise accuracy results, from SESM outperformed EEGNet.

Model explainability

The AOPC scores calculated for each of the 3 data splits are 0.66, 0.61 and 0.70 respectively. The highest observed AOPC score was demonstrated in split 3, which indicates that the most relevant heads share great significance to the model's prediction with minimal redundancy. The results are consistent across data splits, with a standard devition of *0.043*.

In addition to the use of AOPC for an objective measure of explainability, correctly predicted test samples from subject 2 were selected at random. Figure 4 visualises two randomly sampled instances of the W class, selecting the top three most relevant heads. In both samples, heads 1, 3, and 4 were determined to be the most relevant. In Fig. 4a heads 1 and 3 showed the highest relevance, with strong confidence for the W class. Head 1 was the most relevant selecting short sequences of variability, with slight overlap with the sections selected by head 3. Head three selected segments of the signal with high amplitude change with high-frequency variability. Interestingly, head 4 selected different sections of the sample compared to heads 1 and 3 predicting REM as the

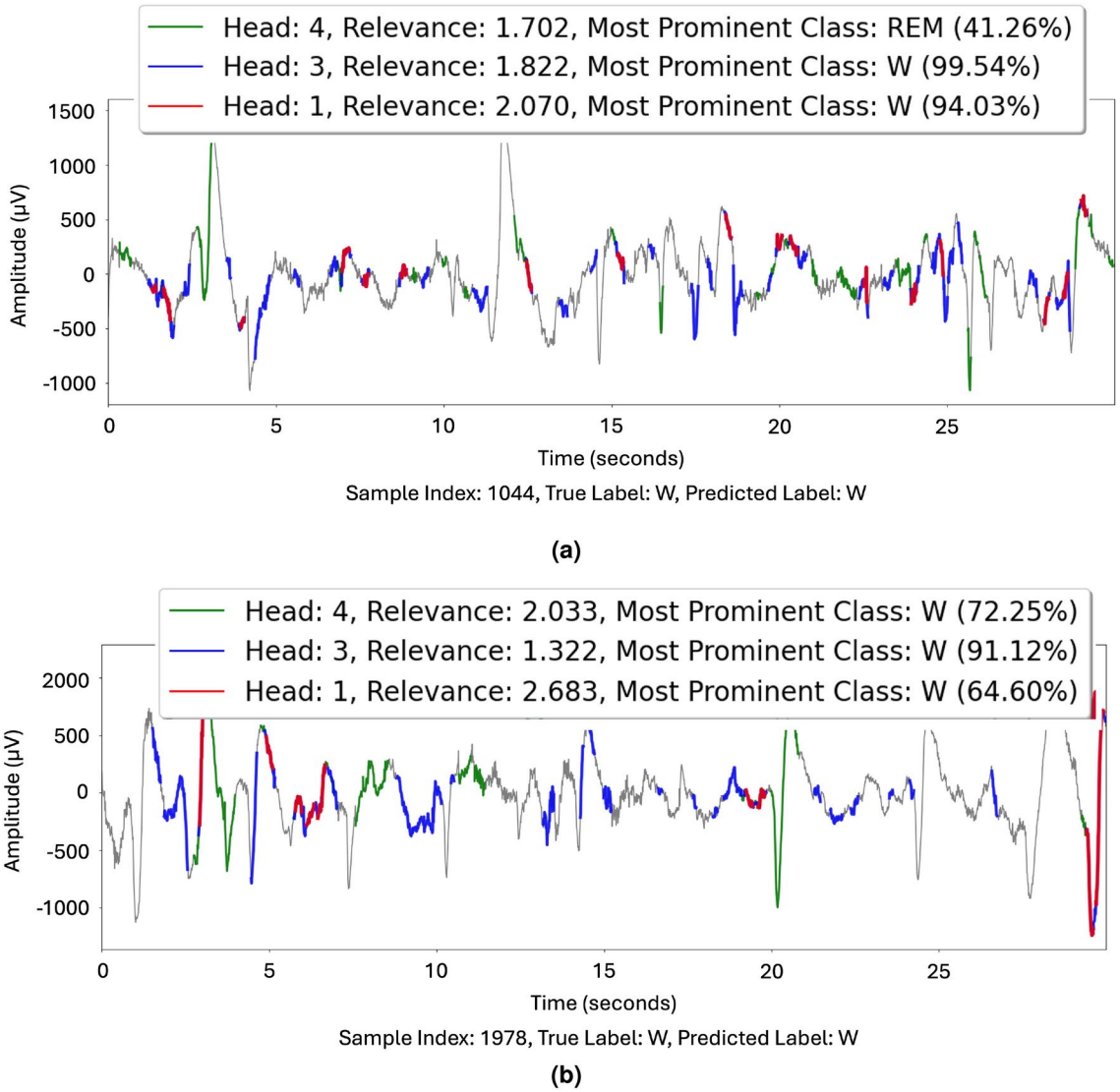


Figure 4. Visualisation of two randomly selected samples from subject 3 for the W class, overlaid with top 3 relevant heads including (**a**) sample 1044 and (**b**) sample 1978.

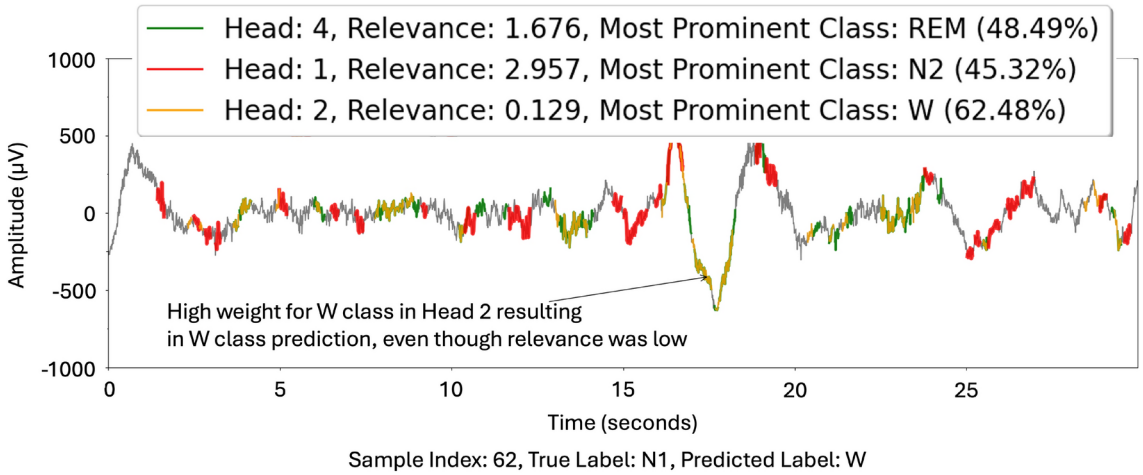


Figure 5. Visualisation of a randomly selected sample where N1 sleep was incorrectly classified as W, overlaid with top 2 relevant heads and head with highest confidence for W class.

Misclassified as	Number of samples	Mean relevance weights of misclassifying heads	Mean relevance weights of N1 classifying heads	T-test (BH corrected p-value)
W	16	1.467	0.517	4.088 (**1.78**$\times 10^{-4}$)
N2	109	1.457	0.439	10.255 (**2.84**$\times 10^{-21}$)
N3	7	1.477	0.438	2.311 (**3.29**$\times 10^{-2}$)
REM	66	1.749	0.495	10.495 (**2.84**$\times 10^{-21}$)

Table 4. Mean relevance weights for heads which provided the incorrect class prediction as the most prominent compared to heads which provided the correct N1 prediction in all samples where N1 was misclassified. The p-values have been corrected for multiple tests using the Benjamini-Hochberg false discovery rate procedure. Significant values are in bold.

most likely class. However, the confidence for the REM class (41.26%) is lower than the confidence for heads 1 and 3 for the W class (94.03% and 99.54% respectively). In Fig. 4b all three of the most relevant heads predicted the W class as the most prominent class. Head 3 selected sections of signal with high amplitude variability and presented the highest confidence for the W class, but had the lowest relevance. In contrast, head 1 showed the highest relevance but the lowest confidence. Head one generated fewer prototypical parts, and overlapped significantly with head 3 as in Fig. 4a. Head 4 selected similar spikes in amplitude for both samples, yet attributed them to REM in the first sample while attributing them to the W class in the second. There was minimal overlap between the prototypical parts generated from head 4 and those generated by heads 1 and 3.

Figure 3a indicates that N1 samples were largely misclassified as either W, N2, or REM, specifically in subject 3 N1 samples were largely misclassified as N2 or REM. To examine the underlying mechanism behind this miss-classification, Fig. 5 visualises a randomly selected sample where N1 sleep was incorrectly classified as W. In this example, the two most relevant heads (1 and 3) predicted N2 and REM sleep with low confidence. Although head 2 exhibited relatively low relevance, it displayed strong confidence in predicting the W class with prototypical W segments. For instance, a prototypical segment generated by 2 illustrates a long segment of significant amplitude variation with a low-frequency component, characteristic of the W class. This sample demonstrates how, despite head 2's lower relevance weight, its high confidence in the W class led to an overall prediction of W. It suggests that a prototypical segment with sufficiently high confidence for a specific class can override concepts deemed more relevant by other heads. Empirically, however, this phenomenon is not the most likely cause of the miss-classification of N1 sleep. Table 4 explores samples where the class was miss-classified as not being N1 and shows the average relevance weights for heads that incorrectly predicted each class as the most prominent compared to heads that correctly predicted N1 in these samples which were incorrectly classified. As shown, the average relevance weights for heads that predicted the wrong class are statistically higher than those that predicted N1. This suggests that the model, in general, makes an incorrect prediction of N1 because heads that predict the wrong class have a misinformed higher relevance than heads that correctly guessed N1 for that sample.

Figure 6 visualises a randomly selected correctly guessed sample from the N2 class, comparing the expected slow-wave and spindle features with the selections from the top three most relevant heads. In this instance, all three heads predicted N2 as the primary class, albeit with low confidence percentages ranging from 60.11 to 69.25%. Heads 1 and 3 shared considerable overlap in their selections, while head 4 predominantly chose distinct segments. In combination, these heads successfully identified two sleep spindles in the signal, along with short

sections depicting slow waves. Irrelevant segments selected by all heads were also present in the middle of the sample. For this portion of the signal, the prototypical parts identified by the heads did not align with sleep spindles or slow waves. This example demonstrates that although the expected features are correctly identified and attributed to the N2 class, the model does not always provide the most concise explanation.

Ablation studies

Below are the findings while ablation studies were carried out to observe the effects of kernel size, the number of heads and diversity thresholds:

1. Table 7a indicate that the accuracy and f1-score increases as the kernel size increases to 30, and steadily declines for $k > 30$. Similarly, the AOPC increases dramatically until $k = 40$ where it decreases and remains consistent.
2. Figure 7b shows the accuracy and AOPC as the number of heads increases for all three of the hold-out test subjects. The accuracy did not change as H increased from 4 to 5, however decreased for $H > 5$, where the number of heads appeared to have little impact on the test accuracy of the model but the F1-score gradually increased.
3. Figure 7c shows the performance Accuracy and AOPC across all hold-out test subjects as the diversity threshold d_{min} increases. The value of d_{min} has no noticeable effect on the accuracy on the test dataset, apart from a slight decrease for $d_{min} = 5$. However, an increase in d_{min} while keeping the number of heads constant demonstrated a steady increase in AOPC until $d_{min} = 9$, after which the AOPC decreased dramatically and stayed relatively low.

Discussion

EEG-SESM performed comparably EEGNet for all three splits, indicating the model has similar predictive power than this implementation of EEGNet. The mean precision and recall across the three data splits for EEG-SESM were not significantly different to EEGNet, however EEG-SESM did demonstrate a statistically significant improvement of mean accuracy across the data splits. These performance results provide evidence that a direct application of SESM, without future modifications, performs similar to EEGNet in this context. Future research is suggested to incrementally improve the performance through modifications discussed in more detail throughout this section.

The standard deviation of the accuracy for EEG-SESM was higher than that of EEGNet, which may suggest some difficulties generalising across subjects compared to EEGNet. Variation in results across subjects is a known issue faced by all EEG studies and is magnified by the use of unseen hold-out subjects in test data[41]. Previous studies have introduced transfer methods for CNN-based architectures used for EEG analysis such as a stage-training strategy[42]. This has been shown to increase the overall accuracy of EEGNet and may be employed in the embedder, conceptizer, and parameterizer encoders of SESM for increased robustness to inter-subject variability. Another recent study using a different prototype-based method on EEG data includes a prototype

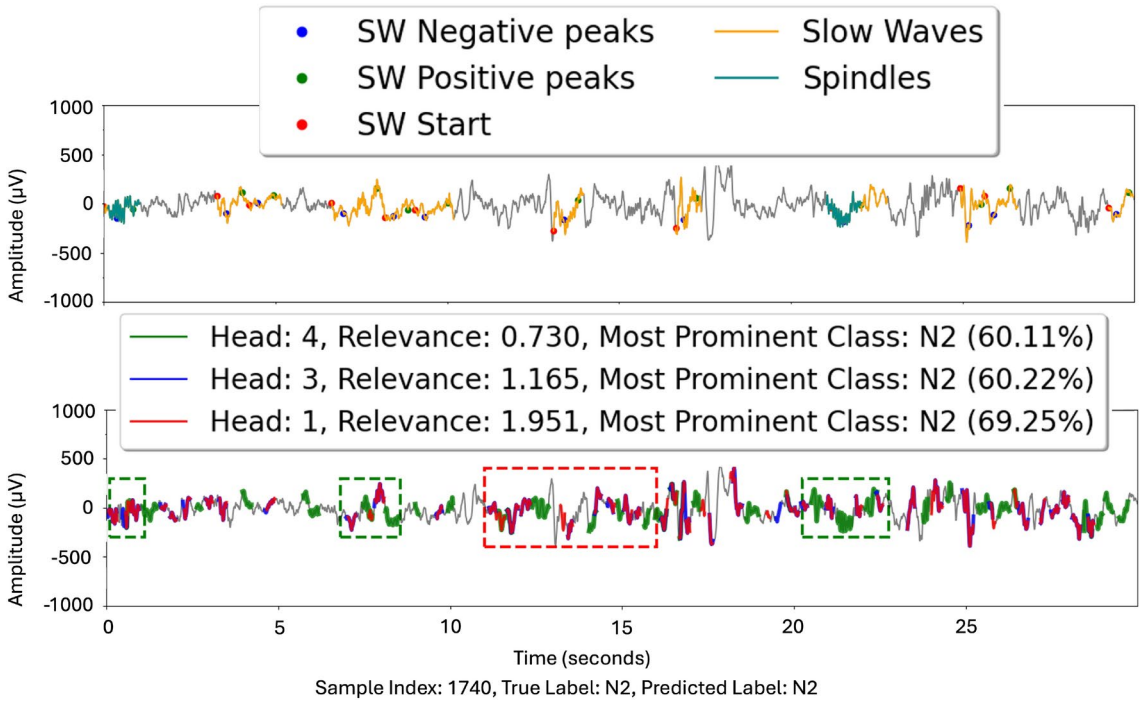
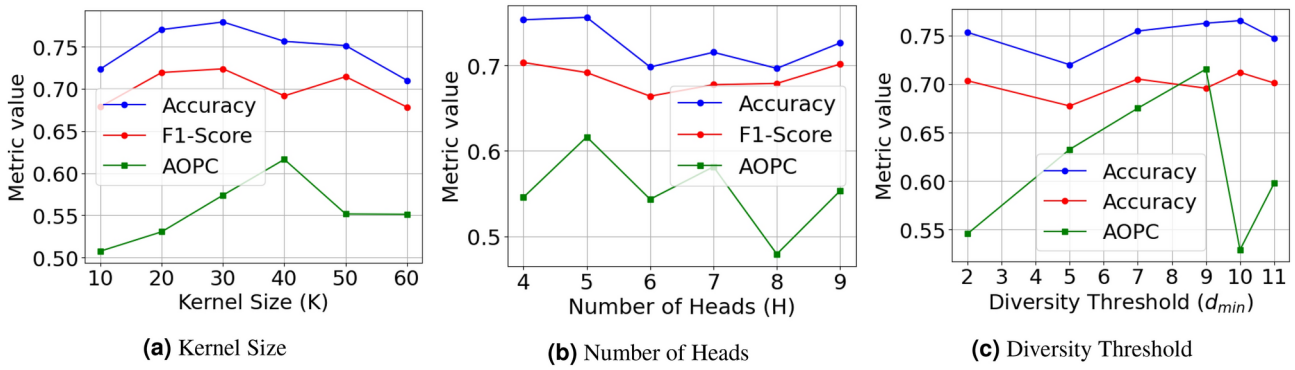


Figure 6. Visualisation of a randomly selected correctly predicted N2 class sample, overlaid with top 3 relevant heads (bottom row) compared to expected slow-wave and spindle features (top row).



(a) Kernel Size **(b)** Number of Heads **(c)** Diversity Threshold

Figure 7. Accuracy (blue), F1-Score (red), and AOPC (green) across all hold-out test subjects for different ablation studies. Like accuracy and the f1-score, the maximum value for AOPC is 1, a higher value represents less redundancy in explanations.

calibration step in which within-dataset cross-subject and cross-dataset samples are used during model testing to improve results[28]. This method is not as easily transferable to the self-attention-based prototype generation of SESM but may be explored in future work.

For both EEG-SESM and EEGNet, the highest accuracy was observed for the N3 class. Conversely, the lowest accuracy was recorded for N1 sleep, which often only involves subtle changes in EEG signals compared to restful wakefulness[31]. Additionally, N1 sleep is the most under-represented class in the dataset. In most cases it was observed that N1 often was misclassified by EEG-SESM as W, N2, or REM, indicating that the concepts learnt do not adequately distinguish the subtle features of N1 sleep. This pattern of N1-REM and N1-N2 miss-classification is common among other sleep classification methods based on single-channel EEG data[43,44]. The prototypical parts offered by the self-attention heads provide insight into the mechanism for the low N1 accuracy for EEG-SESM. In most cases, an incorrect prediction for N1 was because heads that incorrectly predicted the wrong class were of high relevance compared to the heads that correctly predicted N1. A miss-classification due to inappropriate relevance for a sample is likely to be influenced by the paramaterizer network which outputs the relevance weights based on the embedded signal. This suggests that improving the embedder or paramterizer encoders may improve performance.

The motivation for EEG-SESM is to provide a model architecture that is task-independent and can learn concepts regardless of context, and this is the reason for comparing it with a comparably generic EEG-based model such as EEGNet. Previous models specifically designed for sleep stage classification have been presented in previous studies which have performed better than EEG-SESM in their experiments[43,44]. When compared to other studies, EEG-SESM performs most closely with other 1D-CNN-based models, exhibiting a slightly better accuracy on the SleepEDF dataset in one instance[45]. A model that uses a similar attention-based mechanism, albeit for temporary context encoding rather than prototypical-part generation, performed better on the SleepEDF dataset[46]. One critical difference between this architecture and EEG-SESM is the process of feature extraction or embedding. In cases like this where spectral features were explicitly extracted through multi-scale or multi-resolution CNN layers the accuracy was higher than EEG-SESM[44,46]. Important to note though is that the benefit SESM has over these black-box approaches is its intrinsic explainability through the generated prototypical parts. Similar approaches to include multi-scale CNN layers have been included in other prototype-based networks which suggests that the addition of this mechanism to SESM in the embedder or paramterizer may dramatically improve the performance[25,29]. From the class-wise accuracy results and visual explainability analysis, the integration of spectral feature extraction is a highly recommended avenue for future iterations of SESM to be applied to EEG.

The AOPC was relatively consistent across splits and was higher than observed when the SESM architecture was applied to ECG data[26]. There also appears to be no strong correlation between the accuracy of the model and AOPC, that is a change in redundancy amongst the heads does not affect the accuracy of the final prediction, providing further evidence that redundant heads do not impact predictive power[26]. Overall the prototypical explanations typically generate a lot of noise by highlighting seemingly irrelevant segments of the signal, limiting their usefulness in the current context. The best representation of an explanation generated by EEG-SESM was shown for the W class, which identified both a high-amplitude peak and low-amplitude mixed-frequency activity from three separate heads. Prototypical parts from heads appeared to overlap, however, suggesting that although the AOPC is high there still exists redundancy between the heads. Similar useful explanations were common for N2 class predictions, particularly for alpha-spindles which are easily identifiable in the time domain. The presence and apparent detection rate of these time domain characteristic features are may be a reason why the N2 class-wise accuracy was statistically significantly higher for EEG-SESM compared to EEGNet, though future research is required explore this conjecture quantitatively. Features that contain more frequency-rich information were selected with less accuracy, such as slow waves in N2 and N3 sleep. The inability of the model to identify spectral-based bio-markers may explain the lower accuracy for the N1 and N3 classes, as the EEG features for these classes are strongly related to phenomenon in the frequency domain[31]. To the author's knowledge, this prototype generation and analysis represents the first examination of prototypes sampled from data in real-time

aligning to EEG biomarkers. The exploration of other common sleep features identified in the AASM[30] such as K-complexes and vertex sharp waves was not considered in the current study but may provide greater insight into the interpretation of the generated prototypical parts.

The findings of ablation studies can be summarised as follows:

- The results from the kernel size experiments provide evidence that a larger kernel size has an impact on the accuracy and redundancy of heads. A larger kernel size introduced more parameters and a model with higher accuracy and F1-score until $k = 30$. Traditionally, more parameters in a model increase the model's capacity for prediction but increases the likelihood of overfitting. The decrease in accuracy and F1-score for $k > 30$ may be due to this overfitting phenomenon and potentially could be improved by increased regularisation, such as increasing the dropout before the final layer. A larger kernel size also demonstrated a higher diversity loss throughout training. Previous studies using CNN layers for time-series EEG data typically recommend a kernel size half of the sampling rate to capture frequency information of above 2Hz[14]. With limited scope to extract all available features, this suggests one reason why a model trained with a smaller kernel size had increased redundancy. The choice of kernel size in the current study was modified as a hyper-parameter however there is some evidence to suggest that an appropriate kernel size can be learnt from the data[47].
- In contrast to previous applications of the SESM architecture, the number of heads appeared to impact model accuracy[26]. As the number of heads increased from 5 to 6 the accuracy dropped dramatically. This provides new evidence that suggests additional heads may provide noise to the model predictions and decrease a model's performance. With respect to the impact on AOPC, the addition of excessive prototypes contributing to redundancy in the explanations is a known occurrence in prototype-based models and may explain the overall decrease in AOPC as the number of heads increases[48,49]. The unique pattern of alternating AOPC, that is a slight increase followed by a sharp decrease after the addition of another head may imply an increase in redundancy for excessive heads until there is a sufficient number to represent a new concept. A higher number of heads may be harder for end users to interpret, given research has shown that people often prefer simple explanations even if a more complex explanation is more likely[50,51]. Further work is required to see if this preference is valid for the current context.
- In general, the accuracy and F1-score were not affected by changes to the diversity threshold, whereas the AOPC increased as d_{min} increased until $d_{min} = 9$. Small values for d_{min} resulted in a close to 0 diversity loss early in training, which interrupted minimisation for later epochs, resulting in less diverse heads. Conversely, a high diversity threshold ($d_{min} > 9$) increases the minimum distance for heads in the final model and subsequently also results in less diverse heads. Hence, a maximum value for d_{min} should be chosen which reduces redundancy of attention heads (i.e. not too high) while maintaining the utility of the diversity loss throughout training (i.e. not too low). Further analysis is required to determine if the optimal value for d_{min} impacted by, or can be derived from, the number of heads or length of the signal.Limitations of our study can be summarised as follows:
-
- One limitation of this model is the computational complexity involved in performing the multi-headed self-attention mechanism on a long time series. The final model took approximately 35 hours to train on the specified hardware until the early stopping policy was reached and required a larger GPU memory of at least 8 GB. Previous studies have suggested potential lines of research for improving the efficiency of multi-headed self-attention mechanisms such as utilising shared query and key projection[50,52]. Additionally, the SESM architecture may benefit from using shorter epochs, such as 10 seconds, if being used for sleep stage classification in the future as with other studies[33].
- The current iteration of EEG-SESM only supports single-channel data. The use of single-channel data for CNN-based or attention-based sleep stage classification is common[43,44,46], though the AASM rules do provide guidelines for multiple channels[30,31]. The spatial information provided by the addition of multiple channels would be lost by the convolutional layers of the embedder. As such, the selective attention generated from the self-attention mechanism could not highlight sub-regions of channels independently. EEGNet[14] introduced channel-independent processing through the use of depth-wise convolutions in temporal filters, it is possible a similar approach could be taken in the embedder and conceptizer of EEG-SESM to preserve channel independence while generating the concepts.
- Unlike traditional prototype-based methods, the SESM architecture has no direct mechanism for domain experts to alter the learnt concepts to better match expected bio-markers. For example, ProSeNet[24] provides the ability to manually refine prototypes after generation. SCN_{PRO}[25] allows for the indirect manipulation of prototypes through the filter resolution.
- Finally, the interpretability of the prototypical part explanations was only investigated qualitatively and through the AOPC measurement. Different methods for evaluating an explanation of an AI prediction have been suggested, including human evaluation of interpretability via a selection of questionnaires provided to relevant domain experts to quantify a subjective measure of explainability[5,24].

Conclusion

The current work represents the first comprehensive application of a self-attention based prototype method to EEG data, including the qualitative validation of concepts from the sleep stage classification perspective. The findings suggest that though a reasonable accuracy can be achieved for a subset of subjects, there is still inter-subject variability in the results. Objective and qualitative analysis of the prototypical parts offer some insights into the explanations generated from the model, though the alignment to bio-markers is limited by the model's ability to only express features in the time-domain. This limitation may prove the SESM unsuitable for sleep-stage classification where the sleep phenomena demands analysis of EEG signal's frequency components, though

the results are promising to explore applications to other EEG tasks. Based on the results from the current study, the authors recommend future work in adapting SESM to support the integration of spectral information when applying the model architecture EEG. The behaviour of SESM components such as kernel length, diversity threshold, and number of attention heads has also been studied as a useful guide to these future works.

Data availibility

The dataset used in the current research was obtained from PhysioNet[37] in accordance with the Open Data Commons Attribution License (ODC-By) v1.0. The data is available through the following link https://www.physionet.org/content/sleep-edf/1.0.0/.

Received: 1 July 2024; Accepted: 5 November 2024
Published online: 11 November 2024

References

1. Siuly, S., Li, Y. & Zhang, Y. EEG signal analysis and classification. *Health Inf. Sci.*[SPACE]https://doi.org/10.1007/978-3-319-47653-7 (2016) (**MAG ID: 2562547616 S2ID: 9c235554ac1ee5d23621ab4848bd23c53b6ba49c**).
2. Zhao, D., Tang, F., Si, B. & Feng, X. Learning joint space-time-frequency features for EEG decoding on small labeled data. *Neural Netw.* **114**, 67–77. https://doi.org/10.1016/j.neunet.2019.02.009 (2019) (**Tex.eprint: 30897519 tex.eprinttype: pmid**).
3. Amann, J., Blasimme, A., Vayena, E., Frey, D. & Madai, V. I. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 310–310. https://doi.org/10.1186/s12911-020-01332-6 (2020) (**Tex.eprint: 33256715 tex.eprinttype: pmid tex.pmcid: 7706019**).
4. Elshawi, R., Sherif, Y., Al-Mallah, M. H. & Sakr, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Comput. Intell.* **37**, 1633–1650. https://doi.org/10.1111/coin.12410 (2021).
5. Vilone, G. & Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inform. Fusion* **76**, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009 (2021).
6. Arrieta, A. B. et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* **58**, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 (2020).
7. Ravindran, A. S. & Contreras-Vidal, J. L. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Sci. Rep.* **13**. https://doi.org/10.1038/s41598-023-43871-8 (2023).
8. Farahat, A., Reichert, C., Reichert, C., Sweeney-Reed, C. M. & Hinrichs, H. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *bioRxiv* 614784. https://doi.org/10.1101/614784 (2019).
9. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* **10**, 0130140. https://doi.org/10.1371/journal.pone.0130140 (2015).
10. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929. https://doi.org/10.1109/cvpr.2016.319 (2016). arXiv:1512.04150
11. Sturm, I., Bach, S., Samek, W. & Müller, K.-R. Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* **274**, 141–145. https://doi.org/10.1016/j.jneumeth.2016.10.008 (2016) (**Tex.eprint: 27746229 tex.eprinttype: pmid**).
12. Cui, J., Lan, Z., Sourina, O. & Muller-Wittig, W. EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. *IEEE Trans. Neural Netw. Learn. Syst.*[SPACE]https://doi.org/10.1109/tnnls.2022.3147208 (2022). Tex.eprint: 35171778 tex.eprinttype:pmid.
13. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*. 3145–3153 (2017). arXiv:1704.02685.
14. Lawhern, V. J. *et al.* EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. *J. Neural Eng.* **15**, 056013. https://doi.org/10.1088/1741-2552/aace8c (2018). Tex.eprint: 29932424 tex.eprinttype:pmid.
15. Fong, R. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2017.371 (2017). arXiv: 1704.03296.
16. Ellis, C. A., Sattiraju, A., Miller, R. & Calhoun, V. Examining effects of schizophrenia on EEG with explainable deep learning models. In *IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. 301–304. https://doi.org/10.1109/bibe55377.2022.00068 (2022).
17. Wang, H., Zhu, X., Chen, T., Li, C. & Song, L. Rethinking saliency map: A context-aware perturbation method to explain EEG-based deep learning model. *IEEE Trans. Biomed. Eng.*[SPACE]https://doi.org/10.1109/tbme.2022.3218116 (2022). Tex.eprint: 36315542 tex.eprinttype: pmid.
18. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019) (publisher: Nature Publishing Group UK London).
19. Habib, A., Karmakar, C. & Yearwood, J. Interpretability and optimisation of convolutional neural networks based on sinc-convolution. *IEEE J. Biomed. Health Inform.* **27**, 1758–1769 (2022) (publisher: IEEE).
20. Borra, D., Fantozzi, S. & Magosso, E. EEG motor execution decoding via interpretable sinc-convolutional neural networks. In *Mediterranean Conference on Medical and Biological Engineering and Computing*. 1113–1122. https://doi.org/10.1007/978-3-030-31635-8_135 (2019).
21. Borra, D., Fantozzi, S. & Magosso, E. Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination. *Neural Netw.* **129**, 55–74. https://doi.org/10.1016/j.neunet.2020.05.032 (2020) (**Tex.eprint: 32502798 tex.eprinttype: pmid**).
22. Ellis, C. A., Miller, R. L. & Calhoun, V. D. A novel local explainability approach for spectral insight into raw EEG-based deep learning classifiers. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. 1–6. https://doi.org/10.1101/2021.06.10.447983 (IEEE, 2021).
23. Li, O., Liu, H., Chen, C. & Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *AAAI Conference on Artificial Intelligence*. 3530–3537. https://doi.org/10.1609/aaai.v32i1.11771 (2017).
24. Ming, Y., Xu, P., Qu, H. & Ren, L. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 903–913. https://doi.org/10.1145/3292500.3330908 (2019).
25. Ni, J. *et al.* Interpreting convolutional sequence model by learning local prototypes with adaptation regularization. In *International Conference on Information and Knowledge Management*. 1366–1375. https://doi.org/10.1145/3459637.3482355 (2021).
26. Zhang, Y., Neng, G. & Cunqing, M. Learning to select prototypical parts for interpretable sequential data modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6612–6620. https://doi.org/10.48550/arxiv.2212.03396 (2022).
27. Alvarez Melis, D., Alvarez-Melis, D. & Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. *Neural Inf. Process. Syst.* **31**, 7786–7795 (2018).
28. Qiu, L. et al. A novel EEG-based Parkinson's disease detection model using multiscale convolutional prototype networks. *IEEE Trans. Instrum. Meas.*[SPACE]https://doi.org/10.1109/tim.2024.3351248 (2024).

29. Wang, Y. et al. EEG-based emotion recognition with prototype-based data representation. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*[SPACE]https://doi.org/10.1109/embc.2019.8857340 (2019) (**Tex.eprint: 31945990 tex.eprinttype: pmid**).
30. Anderer, P. et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. *Neuropsychobiology* **51**, 115–133. https://doi.org/10.1159/000085205 (2005) (**Tex.eprint: 15838184 tex.eprinttype: pmid**).
31. Iber, C., Ancoli-Israel, S., Chesson, A. L. & Quan, S. F. The American Academy of Sleep Medicine (AASM) manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. (2007).
32. Anderer, P. et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: Validation study of the AASM version of the Somnolyzer 24 × 7. *Neuropsychobiology* **62**, 250–264. https://doi.org/10.1159/000320864 (2010) (**Tex.eprint: 20829636 tex.eprinttype: pmid**).
33. Aboalayon, K. A. I., Faezipour, M., Almuhammadi, W. S. & Moslehpour, S. Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation. *Entropy Int. Interdiscip. J. Entropy Inf. Stud.* **18**, 272 https://doi.org/10.3390/e18090272 (2016).
34. Wilcox, P. *et al.* Diagnostic tests for sleep disorders. In *Pulmonary Function Tests in Clinical Practice*. 217–264 (2009) (publisher: Springer).
35. Supratak, A., Dong, H., Wu, C. & Guo, Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1998–2008. https://doi.org/10.1109/tnsre.2017.2721116 (2017) (**Tex.eprint: 28678710 tex.eprinttype: pmid**).
36. Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C. & Oberye, J. J. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* **47**, 1185–1194. https://doi.org/10.1109/10.867928 (2000) (**Tex.eprint: 11008419 tex.eprinttype: pmid**).
37. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *circulation* **101**, e215–e220 (2000) (publisher: Am Heart Assoc).
38. Supratak, A., & Guo, Y. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2020. 641–644. https://doi.org/10.1109/embc44109.2020.9176741 (2020). Tex.eprint: 33018069 tex.eprinttype: pmid.
39. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. arXiv: Learning (2017).
40. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995) (**MAG ID: 2110065044 S2ID: fcef2258a963f3d3984a486185ddc4349c43aa35**).
41. Kunjan, S. et al. The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis. *BI*, 558–567 https://doi.org/10.1007/978-3-030-86993-9_50(2021).
42. Sameri, J., Zarooshan, H. & Jahed-Motlagh, M. R. A deep transfer learning training strategy for inter-subject classification of EEG signal. *Iran. Conf. Biomed. Eng.*[SPACE]https://doi.org/10.1109/icbme54433.2021.9750313 (2021).
43. Jadhav, P. & Mukhopadhyay, S. Automated sleep stage scoring using time-frequency spectra convolution neural network. *IEEE Trans. Instrum. Meas.* **71**, 1–9. https://doi.org/10.1109/tim.2022.3177747 (2022).
44. Mousavi, S., Afghah, F. & Acharya, U. R. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS ONE* **14**, 1–15. https://doi.org/10.1371/journal.pone.0216456 (2019) (**Tex.eprint: 31063501 tex.eprinttype: pmid tex.pmcid: 6504038**).
45. Tsinalis, O., Matthews, P. M., Guo, Y. & Zafeiriou, S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv: Machine*Learning* (2016).
46. Eldele, E. et al. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 809–818. https://doi.org/10.1109/tnsre.2021.3076234 (2021) (**Tex.eprint: 33909566 tex.eprinttype: pmid**).
47. Tang, W. *et al.* Rethinking 1D-CNN for time series classification: A stronger baseline (2020). Tex.pubstate: preprint.
48. Nauta, M., Jutte, A., Provoost, J. C. & Seifert, C. This looks like that, because ... explaining prototypes for interpretable image recognition. arXiv: Computer*Vis. Pattern Recognit.*[SPACE]https://doi.org/10.1007/978-3-030-93736-2_34 (2020).
49. Sinhamahapatra, P., Heidemann, L., Monnet, M. & Roscher, K. Towards human-interpretable prototypes for visual assessment of image classification models. *VISIGRAPP*[SPACE]https://doi.org/10.5220/0011894900003417 (2023).
50. Cordonnier, J.-B., Loukas, A. & Jaggi, M. *Multi-Head Attention: Collaborate Instead of Concatenate* (2021). Tex.pubstate: preprint.
51. Miller, T., Howe, P. D. L. & Sonenberg, L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv: Artificial*Intelligence* (2017).
52. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI Open*[SPACE]https://doi.org/10.1016/j.aiopen.2022.10.001 (2022).

Author contributions

Brenton Adey has participated in conceptualising the method, running simulations and preparing the manuscript. Ahsan Habib has participated in conceptualising the method, preparing and proof reading the manuscript. Chandan Karmakar has participated in conceptualising the method and proof reading the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.