Automatic Sleep Stage Scoring based on Raw Single-Channel EEG using Transfer Learning

Nirali Parekh, Bhavi Dave, Raj Shah, Kriti Srivastava

Abstract

Medicine has long reached an overwhelming consensus on the importance of sleep in maintaining mental and physiological homeostasis, and the link that sleep disruption has with both disease and mortality. With the advent of the domain of HealthTech, Deep Learning approaches have generated State Of The Art performance in solving several problems in the medicinal arena. The study of sleep- Polysomnography- uses Electroencephalogram (EEG) readings, among other parameters, to gain a clearer picture of a patient's sleep patterns since different brain activities correspond to different stages of sleep. Monitoring and interpreting EEG signals and the body's reactions to the changes in these cycles can help identify disruptions in sleep patterns. Successfully classified sleep patterns can in turn help medical professionals with the prognosis of several pervasive sleep related diseases like Sleep Apnea and the predilection for seizures. To address the pitfalls associated with the traditional manual review of EEG signals that help classify sleep stages, we have trained and analysed the performance of several Convolutional Neural Networks that classify the five sleep stages (Wake, N1, N2, N3, N4 and REM) using data from raw, single channel EEG signals. With the dataset being constant, a comparative analysis of the performance of popular convolutional neural network architectures can serve as a benchmark to the problem of sleep stage classification using EEG signals. We have implemented a CNN to extract time-invariant features and learn stage transition rules among sleep stages from EEG epochs. The models are promising, with some successfully classifying the five stages with a 97% accuracy as well as an f1 score of 0.98.

Background and Motivation

Sleep plays a principal role in mental and physiological homeostasis. Sleep related disorders like insomnia and sleep apnea reduce the quality of life for the multitude of people who are affected. As connections between sleep disruption and both disease and mortality have become more firmly established in *The Extraordinary Importance of Sleep: The Detrimental Effects of Inadequate Sleep on Health and Public Safety Drive an Explosion of Sleep Research* (Susan L. Worley, 2018 Dec)[15], accurate and efficient diagnosis and management of sleep disorders have become increasingly critical. Observing sleep cycles, along with the body's reactions to the changes in these cycles, can help identify disruptions in sleep patterns.

Typically, all-night polysomnographic (PSG) recording consisting of an electrooculogram (EOG), electroencephalogram (EEG) and electromyogram (EMG) are analyzed for determining the quality of sleep. This PSG is divided into segments of 30-second recordings, known as epochs, which are then be classified into different sleep stages by the experts by following the guidelines such as the Rechtschaffen and Kales (R&K) [19] and the American Academy of Sleep Medicine (AASM) [18]. This process is known as sleep stage scoring or sleep stage classification. This procedure is, however, manual, hence labor-intensive and time-consuming. Experts must visually inspect all epochs and label their sleep stages of entire PSG recordings. Thus, automatic sleep scoring for healthcare and well-being is in high demand.

Artificial intelligence (AI) is already delivering on making aspects of health care more efficient, with the potential to support clinical and other applications that result in more insightful and effective care and operations. Deep Learning's efficiency in handling large amounts of data and the power to learn hidden features automatically is what makes it popular in numerous domains, and has hence received considerable attention from the sleep research community. Deep Learning can aid physicians at hospitals and health systems in sleep scoring by providing them with real-time, automatic classification that they can alter and monitor based on their personal expertise. Many sleep scoring methods using Deep Learning for automatic classification of PSG data have been proposed. Some works have been discussed in the section \Lit_review

Literature Review

Various review works have been focussed on the use of EEG waves in the problem of sleep scoring techniques. In *Deep learning-based electroencephalography analysis: a systematic review* (Yannick Roy et al.,14 August 2019 ,Journal Of Neural Engineering) [4] have reviewed 154 papers that apply DL to EEG, published between January 2010 and July 2018, and spanning different application domains such as epilepsy, sleep, brain–computer interfacing, and cognitive and affective monitoring. The review also provides detailed methodological information on the various components of a DL-EEG pipeline to inform their own implementation. A *review of* *automated sleep stage scoring based on physiological signals for the new millennia*(Faust et al., April 2019)[9], has provided a comprehensive review of automated sleep stage scoring systems, since the year 2000. They analyse the system that were developed for Electrocardiogram (ECG), Electroencephalogram (EEG), Electrooculogram (EOG), and a combination of signals.

Mixed Neural Network Approach for Temporal Sleep Stage Classification (Hao Dong et Al.)[5] proposes a Mixed Neural Network (MNN) that simultaneously aims to target the issues of population heterogeneity and temporal pattern recognition. Their novel architecture is composed of a rectifier neural network which is suitable for detecting naturally sparse patterns and a long short-term memory (LSTM) for the detection of temporally sequential patterns. Their model achieved a promising accuracy of 85.92%, outperforming SVM, RF and MLP in a comparative analysis.

In *SLEEPNET: Automated Sleep Staging System via Deep Learning* **(**Siddharth Biswa et Al.) [6] have proposed a deployed annotation tool for sleep staging. SleepNet uses a deep recurrent neural network trained on the largest sleep physiology database assembled to date, consisting of PSGs from over 10,000 patients from the Massachusetts General Hospital (MGH) Sleep Laboratory.The best performing instance of SleepNet uses expert-defined features to represent each 30-sec interval and learns to annotate EEG using a recurrent neural network (RNN).Their model analyses and compares various algorithms like Logistic Regression, Tree Boosting, Multilayer Perceptron, CNN, RNN and RCNN.

In *Mixed Neural Network Approach for Temporal Sleep Stage Classification*(Hao Dong et Al.)[7] have proposed use of a Mixed Neural Network (MNN) to solve both the population heterogeneity and temporal pattern recognition problems. Their MNN is composed of a rectifier neural network which is suitable for detecting naturally sparse patterns, and a long short-term memory (LSTM) for detection of temporally sequential patterns.The proposed Mixed Neural Network and the corresponding training method work well for sleep stages classification problem compared with SVM, RF and MLP.

SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach **(**Sajad Mousavi et Al. ,2019 May 7)[8] have proposed an automatic sleep stage annotation method called SleepEEGNet using a single-channel EEG signal . The authors have applied sequence of deep learning models - CNN for feature extraction, BiRNNs for capturing temporal inf and an attention network to let the model learn the most relevant parts of the input sequence while training .The authors used the synthetic minority over-sampling technique (SMOTE) to tackle class imbalance problems.

A deep learning approach to detect sleep stages (Klara Stuburić et al. , Procedia Computer Science)[10] have presented the implementation of deep learning methods for sleep stage detection by using three signals: heartbeat signal, respiratory signal, and movement signal. They employ two different neural networks for this classification problem: the convolutional neural network (CNN) and the long-short term memory network (LSTM). However their model performs poorly giving only an accuracy of 40% and F1 score of 37% in classification of the 5 stages of sleep.

SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging (Huy Phan et al., 2019 Jan 31) [11] has proposed a hierarchical recurrent neural network. The author treated automatic sleep staging as a sequence-to-sequence classification problem to jointly classify a sequence of multiple epochs at once. They achieve an overall accuracy and F1 score of 87.1% and 83.3% respectively.

Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning (Huy Phan at al) [12] presented a deep transfer learning approach to address the problem of insufficient data in many sleep studies and to improve automatic sleep staging performance on small cohorts. They adopted the MASS dataset as the source domain and three different sleep databases are used as the target domains.

In their paper, *Intra- and Inter-epoch Temporal Context Network (IITNet) Using Sub-epoch Features for Automatic Sleep Scoring on Raw Single-channel EEG* that proposed a novel deep learning model (Hogeon Seo et al.)[13] considered the inter- and intra-epoch temporal contexts using raw single-channel EEG. They model a deep CNN based on a modified ResNet-50 extracts the sleep-related features and the RNN via two-layered BiLSTM learns the transition rules.Their results show that the proposed temporal context learning at both the intra- and inter-epoch levels is effective to classify the time-series inputs.

Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation (Khald Aboalayon et Al.) [16] proposes the development of an Automatic Sleep Stage Classification (ASSC) system for detecting sleep stages using simple statistical statistical features (called EnergySis and Maximum-Minimum Distance) that are applied to 10 s epochs of single-channel EEG signals. Their solution that used Decision Tree achieved an average classification sensitivity, specificity and accuracy of 89.06%, 98.61% and 93.13% on the PhysioNet Sleep European Data Format (EDF) Database. The work is most notable for its focus on making the design ideal for being implemented in an embedded device in a real world setting in real time.

The paper *DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification* (H. Phan et al.) [24] proposes an efficient CNN for sleep stage classification. The simplified CNN is capable of learning features at a multitude of temporal resolutions while capturing time shift-invariance property of EEG signals because of its 1-max pooling layer. The work also creates a novel method to discriminatively learn a frequency-domain filter bank with a deep neural network (DNN) to preprocess the time-frequency image features. The model achieves an accuracy of 82.6% on the popular Sleep-EDF dataset.

Automatic Sleep Stage Classification Using Temporal Convolutional Neural Network and New Data Augmentation Technique from Raw Single-Channel EEG (H. Phan et al.) [25] proposed a framework for sleep stage classification from single-channel EEG using a CNN to initially extract features that could serve as an input to a Temporal Convolutional Neural Network (TCNN). The TCNN would help extract the temporal features from the extracted features vector of the CNN to achieve better results. The framework achieved an accuracy of 85.39% on the sleep EDF dataset.

Table 2.1 : Literature Review Analysis

Ref	Task	Dataset used	Feature type	Architecture	Accuracy

[11]	Sleep Scoring	Montreal archive of sleep studies (MASS) open access dataset	Raw signals	SeqSleepNet. Hierarchical RNN	87.1%
[12]	Sleep Scoring	Sleep-EDF SC and ST	learned	Transfer Learning Using CNN	84.3%
[13]	Sleep Scoring	SleepEDF, Montreal Archive of Sleep Studies (MASS) and SHHS	Sub epochs	IITNet - transfer learning + bidirectional LSTM	86.2%
[6]	Sleep Scoring	Massachusetts General Hospital Sleep Laboratory	Raw signal	SleepNet - RNN	85.76%
[7]	Sleep Scoring	Montreal archive of sleep studies (MASS) open access dataset	handcrafted	SVM	79.7%
[5]	Sleep Scoring	Montreal Archive of Sleep Studies (MASS) and Sleep-EDF	learned	DeepsSleep Net - CNN	80.7%
[16]	Sleep stage Scoring	PhysioNet Sleep European Data Format (EDF) Database	handcrafted	Decision Tree	89.06%

[24]	Sleep stage classification	Sleep-EDF dataset	Learned	CNN + DNN (novel time-frequency image features)	82.6%
[25]	Sleep stage classification	Sleep-EDF dataset	Learned	CNN + Temporal CNN + Conditional Random Field Layer	85.39

Research Gaps

- The review clearly demonstrates that most research uses only single-channel EEG signals to make classifications, and does not utilize all the available signal data including Fpz and Pz. The paper hypotheses that these signals contain valuable information about the temporal and spatial features of the waves and using all the signals can boost the accuracy.
- The existing architectures are also extremely bulky, requiring numerous layers and heavy preprocessing that inevitably increases the time required for training. Transfer learning can be invaluable in such a scenario.

Methodology

A. Dataset Used

In this study, two cohorts from the Physionet Sleep-EDF extended dataset contributed in 2018 are used for experimentation. The dataset contains PSG records and their corresponding sleep stages labeled by human sleep experts. These adopted cohorts have diverging health conditions, i.e. healthy (Sleep-EDF-SC) vs. mild sleep difficulty (Sleep-EDF-ST) [26], [27].

Sleep-EDF-SC: This is the Sleep Cassette (SC) subset of the Sleep-EDF Expanded dataset [26], [27], consisting of 20 subjects aged 25-34. Two subsequent day-night PSG recordings were collected for each subject.

Sleep-EDF-ST: This is the Sleep Telemetry (ST) subset of the Sleep-EDF Expanded dataset [26], [27] which was collected for studying the temazepam effects on sleep. The subset consists of 22 subjects aged 18-79 with mild difficulty falling asleep. The PSG signals recorded after placebo intake are made available. Manual annotation was done similar to the Sleep-EDF-SC subset.

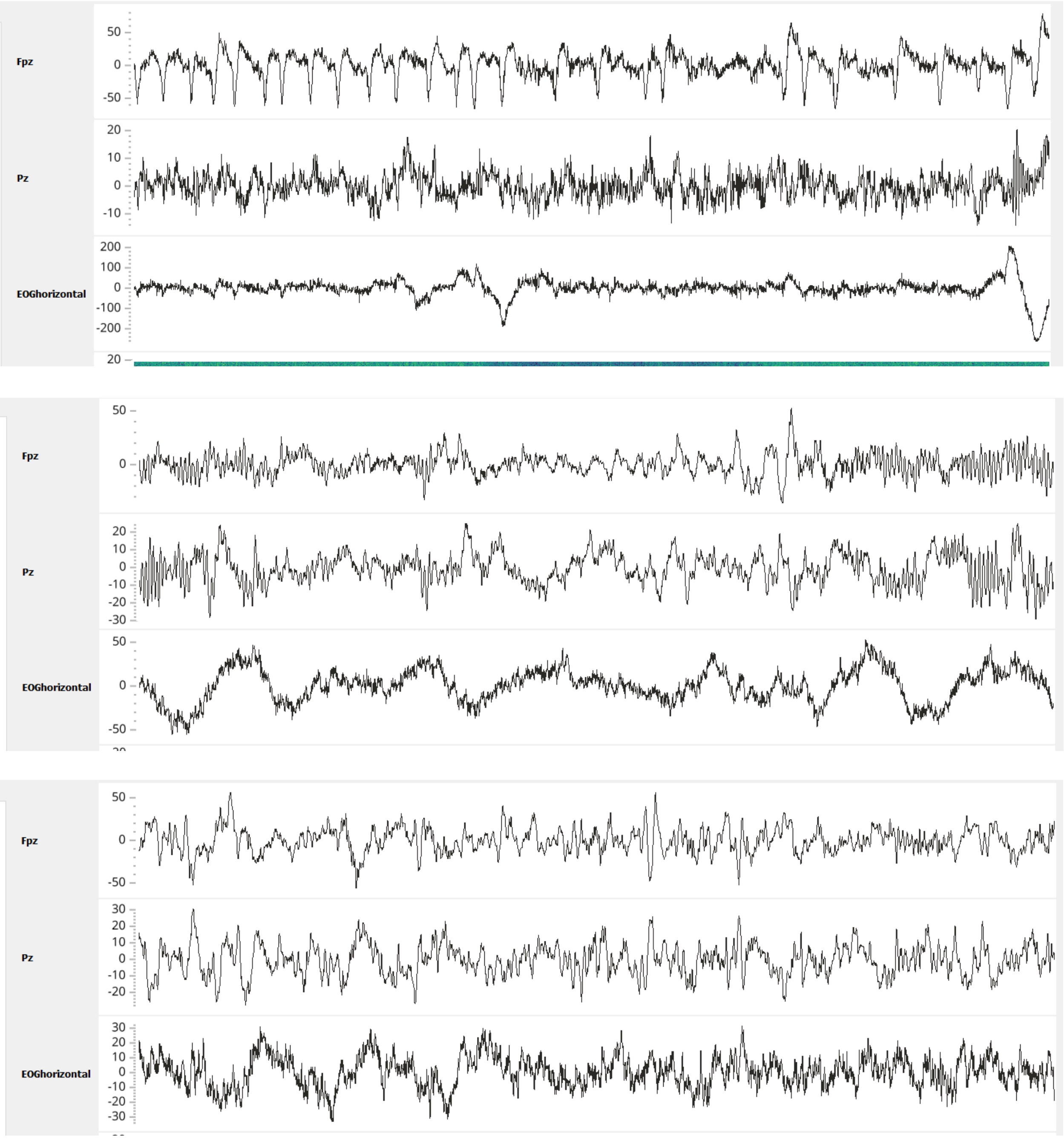
Table 2: The demographic information of the cohorts of Sleep EDF dataset

Dataset	Category	No. of subjects	Age				Average epochs per recording
			mean	min	max	Std dev	
Sleep Cassette	male	77	59.3	26	97	23.31	2650
	female	41	58.5	25	101	21.6	
	total	36	58.9	25	101	22.27	
Sleep Telemetry	male	7	35.71	20	60	17.87	2453
	female	15	50.85	18	79	17.55	
	total	22	40.18	18	79	18.09	

Table 2 summarizes the demographic information of datasets, including gender distributions and age characteristics. In each of the cohorts, each of the 30-second PSG epoch were manually labelled into one of eight categories {W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN} by sleep experts according to the R&K standard [19]. Similar to previous works [23, 24, 5, 25], N3 and N4 stages were merged into a single stage N3 and MOVEMENT and UNKNOWN categories were excluded. To conduct our experiments, we adopted the Fpz-Cz EEG and Pz-Oz EEG channels in this study.

B. Data Exploration

Using the Visbrain tool [20], the visualizations of the sample waveforms of different stages of sleep in an epoch of 30 seconds are shown in Figure 3.



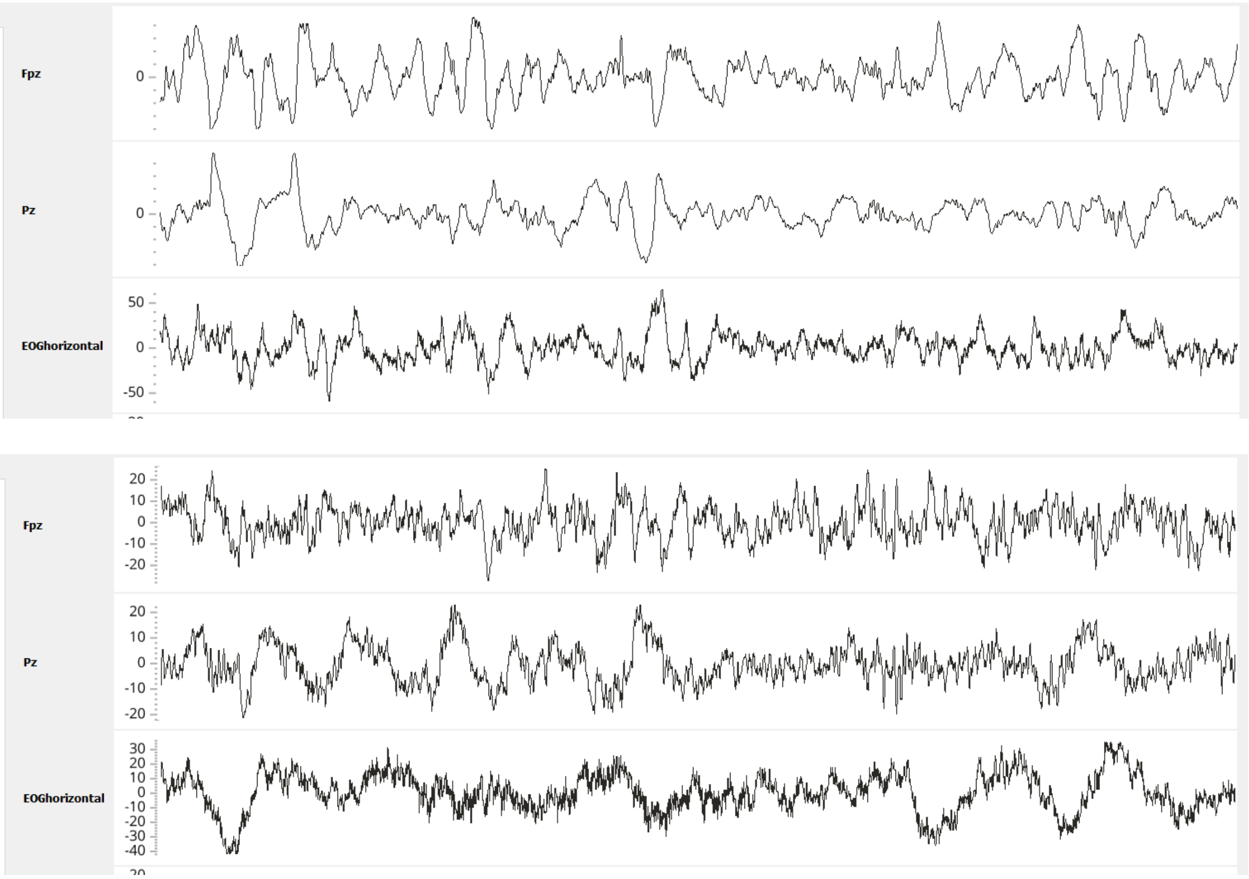


Figure 3 Raw EEG waveforms for each of the sleep stages : (a) Wake stage (b) N1 stage (c) N2 stage (d) N3 stage (e) REM stage

C. Preprocessing

The data set is given as EDF [25] file for each subject. One EDF file represents one PSG record. For the scope of this work, only two signals were needed from the original files: Fpz-Cz EEG, and Pz-Oz EEG signals. By using pyEDFlib library [30], the mentioned signals were extracted and processed. Signals from each subject were divided into time sequences of 3000 timesteps, which, with the frequency of 100 Hz, corresponds to 30 seconds epochs. The tabular dataset obtained for each record is shaped like a three dimensional vector as following: (n epochs, y timesteps, z features) where n samples are the number of rows i.e. epochs inside the data set, n timesteps are the number of timesteps in each sample, 3000 in this case and z features is the number of signals, which is 2 for our study i.e. Fpz-Cz and Pz-Oz. Each data sample that represents one epoch has 480 timesteps, and each timestep has three values, one for every signal.

The obtained dataset was highly unbalanced with a high number of instances i.e. epochs of Wake stage. Due to the uneven distribution of data among the sleep stages, undersampling has been performed to address the data-skewing problem. Near equal number of data points were sampled for the stages Wake, N1, N2, N3 and REM.

Since great results have been achieved by processing data with deep learning techniques like convolutional neural networks (CNN) with images as input. CNN's great performance for reading, processing and extracting the most important features of two dimensional data have highly contributed to its popularity.
In scenarios where input data isn't formatted as an image, like time series, transformation methods have been applied to apply CNN models for our data. In this study, such transformations have been used to convert the EEG signal data to images. So, the sampled epochs obtained in the previous steps are first converted into signal images of grayscale images like the one shown in figure 6.
This preprocessing step helps to detect hidden features and topologies like homogeneity, periodicity, driftnes and dispersion when dealing with time series like EEG signal data.

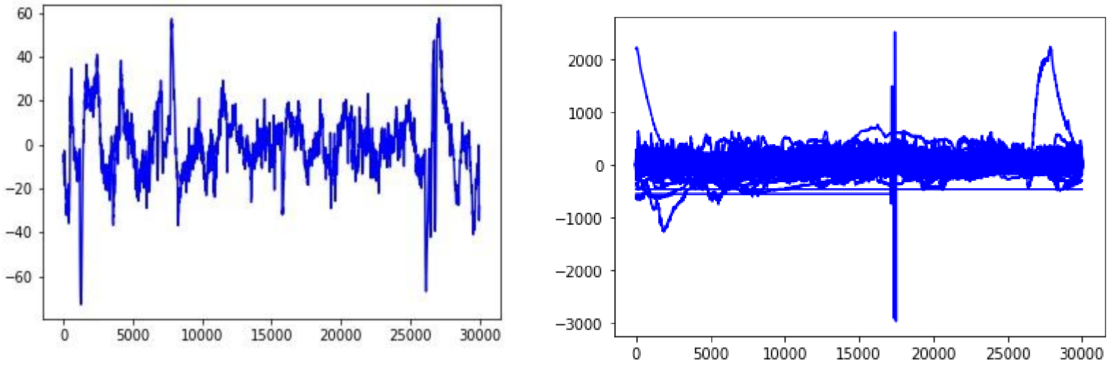


Fig 6. The plotted signals after raw data converted to graphs (a) sample epoch from Sleep Cassette cohort (b) Sample epoch from Sleep Telemetry cohort

Thus, our image data is ready to be fed into the proposed Deep Learning architectures for training the models.

Experimental Architectures

A convolutional neural network is a subtype of deep neural networks which uses alternating layers of convolutions and pooling with trainable filter banks (made up of kernels) per layer. Kernels scan over a receptive field to compute the output feature map with a dot product and bias computation. The result is passed through an activation function, and then subsampled using max pooling to reduce sensitivity to distortions of stimuli in the upper layers. CNNs generally employ the Rectified Linear Unit activation function that converts negative values to 0, extracting only the prominent features. The max pooling layer helps reduce complexity while allowing prominent features to pass through to upper layers of the network . This process is repeated to gain relevant, abstract and location invariant features that are accurate depictions for recognition problems. CNNs are location equivariant in the sense that they preserve the location of features, but extract the same features over the entire image. This manner of feature learning often produces state of the art results. Apart from dominating the domain of computer vision and 11 image analysis, CNNs are also apt for 1D problems like time series, and 3D image classification, because of their reliance on a method's ability to learn features. CNNs greatly reduce the number of parameters when compared to traditional feedforward networks because of their unique design consisting of weight sharing and pooling. Thus, because of sparse-connectivity and weight sharing, we are able to achieve a similar degree of complexity with a significantly smaller number of parameters with a Convolutional layer.

This paper performed experiments on several CNN architectures of varying complexity to analyse the effects of the different design features to the problem statement.The experiments were run with the pretrained models trained on the ImageNet dataset.The last layer to reflect the five sleep stages. With the dataset being constant, a comparative analysis of the performance of popular convolutional neural network architectures can serve as a benchmark to the problem of sleep stage classification using EEG signals. The architectures chosen vary in depth, parameters, design, complexity and emphasis. A comparative analysis of this performance can help give insight to the aspects that will be most relevant to an ideal solution. The following section provides an overview to the key features of the various CNNs trained on the EEG dataset.

AlexNet

AlexNet takes in colour images of dimensions 227x227 that are passed through 5 alternating convolution layers and 3 fully connected layers, with the last fully connected layer having neurons equivalent to the number of classes. The architecture employs max-pooling after convolution layer 1, convolution layer 2 and convolution layer 5. The count of the total number of layers in the architecture does not include the max-pooling layers as they do not carry weights. Thus, AlexNet has 8 layers. AlexNet was proposed in 2012 as a significant improvement on the CNN performance at the time and has been used as a benchmark since. Its key contribution was using the ReLU activation function instead of tanh or sigmoid to achieve faster training times.

VGG

During the design of the VGGNet, it was found that alternating convolution & pooling layers were not required, and consequently VGG uses multiple Convolutional layers in sequence with pooling layers in between. Input of dimensions 224x224 is taken and a kernel size 3x3 is maintained throughout the network with only a change in depth between layers. Appropriate padding is provided to maintain the dimensions across the layers. The VGG architecture has several versions like the VGG-16 (a 16 layer network), VGG-19 (a 19 layer network) and so forth.VGG demonstrated a solution to the pertinent problem of increasing depth of a CNN - the small-size convolution filters in the architecture allowed VGG to have a large number of weight layers;that increased depth and improved performance.

ResNet

As the depth of the VGG network was increased, the model performance began to deteriorate. It was hypothesised that the gradients were not able to flow well through the deeper network. The scientists found that the information from the input was getting highly morphed as it reached the deeper output layers, and consequently proposed a solution of passing the input information repeatedly in stages creating a Residual Network or the ResNet. In ResNet after every two layers the input given to the first layer along with the output obtained at the second layer. The information from the input that was getting highly morphed and by the time it reached the output layers is now passed as a residue of the input once again with the output. This helped the gradients to flow back better, improving training. The general form of all the stacked “Residual Units” comprising a deep residual network is-

$$y \quad = h(x \quad) + F(x \quad, W \quad)$$
$$x_{\quad +1} = f(y \quad)$$

where x and $x_{\quad +1}$ are input and output of the $l{-}th$ unit, and F is a residual function. $h(x) = x$ is an identity mapping and f is the ReLU function. ResNet’s residual connections between layers which were invaluable training models as deep as 151 layers.

DenseNet

DenseNets connect every layer to every other layer, implying that for L layers, there are L(L+1)/2 direct connections. Every layer uses the feature maps of all the preceding layers as inputs, and consequently its output feature maps are used as input for subsequent layers. The network is divided into densely connected blocks within which the feature map size remains the same. This facilitates both downsampling and feature concatenation. DenseNets require fewer parameters than a comparable size traditional CNN as the need to learn redundant feature maps is eliminated. It also facilitates the flow of information and gradient as each layer has direct access to the input and the gradient of the loss function. The architecture consists of four DenseBlocks with varying number of layers, with the convolution operations inside each of the architectures as the Bottle Neck layers. In other words, the 1×1 convolutions are introduced as a bottleneck layer before each 3×3 convolution to reduce the number of input feature-maps, consequently improving computational efficiency.

SqueezeNet

SqueezeNets attempt to reduce the number of parameters in the network by using Fire modules which have a squeeze layer of 1x1 convolutions that can decrease parameters by restricting the number of input channels in every layer. SqueezeNets are therefore very low latency. They achieve AlexNet level accuracies while having ~50 million fewer parameters. A Fire module consists of a squeeze convolution layer that feeds into an expand layer that has a mix of 1×1 and 3×3 convolution filters. SqueezeNet starts with a convolution layer, followed by 8 Fire modules, and concludes with a final convolution layer. The number of filters per fire module decreases with depth. Inspired by ResNet, the architecture also employs bypass.

MobileNet

MobileNets utilise Depth convolutions and point convolutions, reducing the comparison and recognition time. They are designed to increase accuracy while simultaneously factoring the resource constrained environments of embedded devices or mobile phones, therefore getting their name. They also reduce the number of parameters and hence latency. To construct an even smaller and computationally cheap model, MobileNets also have useful model-shrinking parameters that help make the network thinner in a uniform manner at every layer.The architecture helps explore a reasonable accuracy, latency and size trade off. MobileNet has 28 layers if the depthwise and pointwise convolutions are considered to be distinct. All layers in mobilenet are built on depth wise separable convolutions followed by batch normalization and ReLU non-linearity, except for the first layer that is a fully convolutional layer.

Results

A. Evaluation measures

After training, the model has been evaluated on the test set that it has never seen before. The model was trained on 80% of the data and tested on 20% of the complete data set.

1. Accuracy: It is the fraction of classifications the model got correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy is a widespread measurement for evaluation but is not always reliable. In cases of imbalanced classes where one class is significantly more represented, the model will give high accuracy by classifying every sample into that class. That is why precision and recall are introduced because they are giving a better insight into classification performance.

2. Precision: Proportion of the positive identifications which are correct.

$$Precision = \frac{TP}{TP + FP}$$

where, TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative

3. Recall: Proportion of actual positives which were identified correctly.

$$Recall = \frac{TP}{TP + FN}$$

F1 score should be taken as a reliable metric as it is not influenced by bias in the number of samples per class.

4. F1-score: It is a weighted average of Precision and Recall. It is calculated as the harmonic mean of precision and recall.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

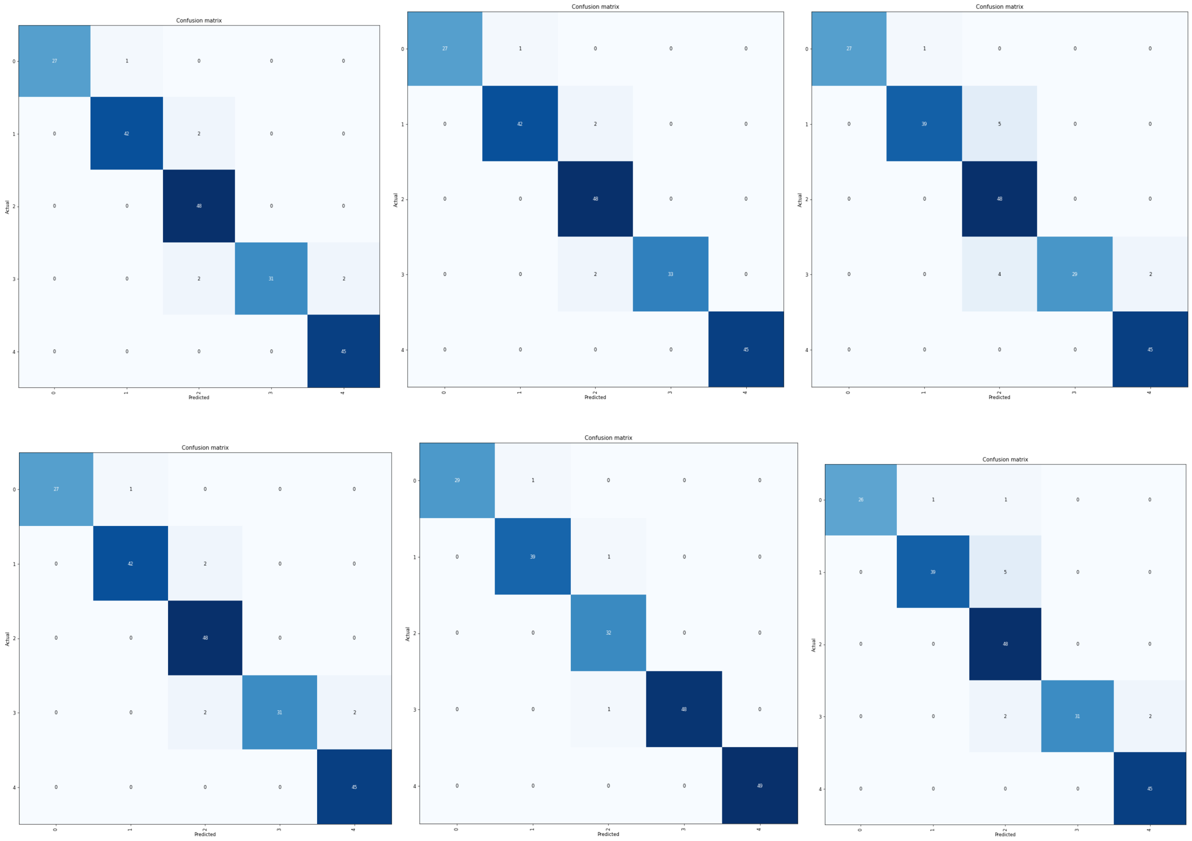
B. Results

The models were used to classify the data into 5 classes - the Wake stage, N1 stage, N2 stage, N3 stage and REM stage. The results are shown in Table 3, which shows previously described measures. All measurements were calculated on test data. For every model and for each classification, learning accuracy reached around 94-95%. Comparing the models, for the Sleep EDF -SC cohort, the best performing model for Fpz-Cz is VGG-19 and for Pz-Oz, it is MobileNet-v2. Similarly, for the Sleep EDF- ST cohort, MobileNet-v2 shows the best performance for Fpz-Cz and Pz-Oz channels both.

Dataset	Channel	Metric	Model					
			ResNet-50	MobileNet-v2	AlexNet	DenseNet121	VGG-19	SqueezeNet
Sleep EDF Sleep Cassette (SC)	Fpz-Cz	Accuracy	0.96	0.97	0.94	0.96	0.98	0.94
		Precision	0.97	0.98	0.95	0.97	0.98	0.96
		Recall	0.96	0.97	0.94	0.96	0.98	0.94
		F1- score	0.96	0.98	0.94	0.96	0.98	0.95
	Pz-Oz	Accuracy	0.98	0.99	0.98	0.97	0.98	0.96
		Precision	0.99	1.00	0.99	0.97	0.9	0.97
		Recall	0.98	0.99	0.98	0.98	0.99	0.96
		F1- score	0.99	0.99	0.98	0.97	0.99	0.96
Sleep EDF Sleep Telemetry (ST)	Fpz-Cz	Accuracy	0.96	0.97	0.95	0.95	0.96	0.95
		Precision	0.96	0.98	0.95	0.95	0.95	0.95
		Recall	0.95	0.97	0.95	0.96	0.96	0.95
		F1- score	0.95	0.97	0.95	0.95	0.95	0.95
	Pz-Oz	Accuracy	0.94	0.95	0.95	0.94	0.94	0.91
		Precision	0.94	0.96	0.95	0.95	0.95	0.93

		Recall	0.94	0.95	0.95	0.94	0.93	0.91
		F1- score	0.94	0.95	0.95	0.94	0.94	0.91

Figure 5 and 6 presents the confusion matrices achieved by each model using the Fpz-Cz and Pz-Oz channels respectively of the Sleep-EDF- Sleep Cassette Cohort. Similarly, Figure 7 and 8 shows the confusion matrices using the Fpz and Pz-Oz channels respectively of the Sleep EDF Sleep Telemetry cohort. The main diagonals in each confusion matrix denote the true positive (TP) values which indicate the number of stages scored correctly. It can be seen from the tables (the confusion matrices' parts) that TP values are significantly higher than other values in the same rows and columns, indicating notable performance of the models.



	0	1	2	3	4
0	27	1	0	0	0
1	0	42	2	0	0
2	0	0	48	0	0
3	0	0	2	31	2
4	0	0	0	0	45

	0	1	2	3	4
0	27	1	0	0	0
1	0	42	2	0	0
2	0	0	48	0	0
3	0	0	2	33	0
4	0	0	0	0	45

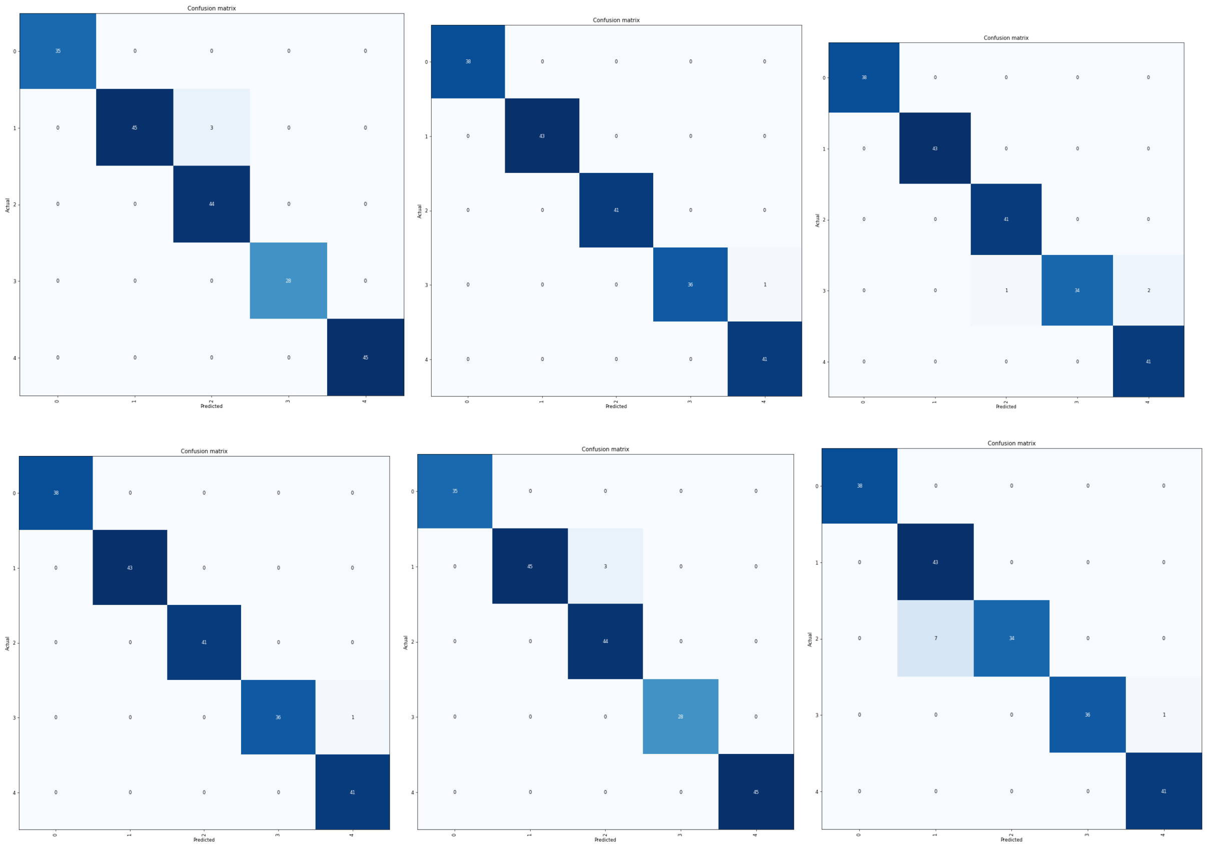
	0	1	2	3	4
0	27	1	0	0	0
1	0	39	5	0	0
2	0	0	48	0	0
3	0	0	4	29	2
4	0	0	0	0	45

	0	1	2	3	4
0	27	1	0	0	0
1	0	42	2	0	0
2	0	0	48	0	0
3	0	0	2	31	2
4	0	0	0	0	45

	0	1	2	3	4
0	29	1	0	0	0
1	0	39	1	0	0
2	0	0	32	0	0
3	0	0	1	48	0
4	0	0	0	0	49

	0	1	2	3	4
0	26	1	1	0	0
1	0	39	5	0	0
2	0	0	48	0	0
3	0	0	2	31	2
4	0	0	0	0	45

Fig 5. Confusion Matrices for different models using Fpz-Cz channel of Sleep-EDF -Sleep Cassette dataset - (a) ResNet50, (b)MobileNet-v2, (c) AlexNet (d) DenseNet (e) VGG-19 (f) SqueezeNet



Confusion matrix

Actual	0	1	2	3	4
0	35	0	0	0	0
1	0	45	3	0	0
2	0	0	44	0	0
3	0	0	0	28	0
4	0	0	0	0	45

Predicted

Confusion matrix

Actual	0	1	2	3	4
0	38	0	0	0	0
1	0	43	0	0	0
2	0	0	41	0	0
3	0	0	0	36	1
4	0	0	0	0	41

Predicted

Confusion matrix

Actual	0	1	2	3	4
0	38	0	0	0	0
1	0	43	0	0	0
2	0	0	41	0	0
3	0	0	1	34	2
4	0	0	0	0	41

Predicted

Confusion matrix

Actual	0	1	2	3	4
0	38	0	0	0	0
1	0	43	0	0	0
2	0	0	41	0	0
3	0	0	0	36	1
4	0	0	0	0	41

Predicted

Confusion matrix

Actual	0	1	2	3	4
0	35	0	0	0	0
1	0	45	3	0	0
2	0	0	44	0	0
3	0	0	0	28	0
4	0	0	0	0	45

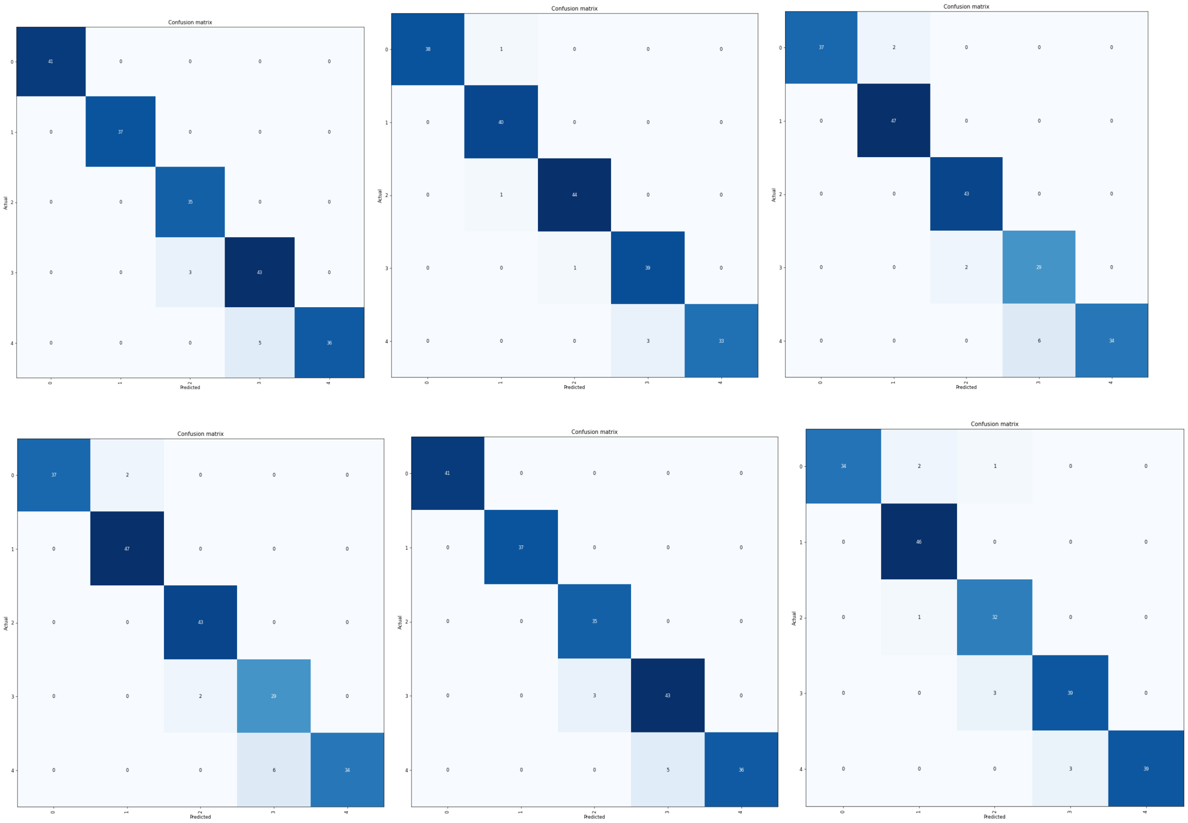
Predicted

Confusion matrix

Actual	0	1	2	3	4
0	38	0	0	0	0
1	0	43	0	0	0
2	0	7	34	0	0
3	0	0	0	36	1
4	0	0	0	0	41

Predicted

Fig 6.Confusion Matrices for different models using Pz-Oz channel of Sleep-EDF - Sleep Cassette dataset - (a) ResNet50, (b) MobileNet-v2, (c) AlexNet (d) DenseNet (e)VGG-19 (f) SqueezeNet



	0	1	2	3	4
0	41	0	0	0	0
1	0	37	0	0	0
2	0	0	35	0	0
3	0	0	3	43	0
4	0	0	0	5	36

	0	1	2	3	4
0	38	1	0	0	0
1	0	40	0	0	0
2	0	1	44	0	0
3	0	0	1	39	0
4	0	0	0	3	33

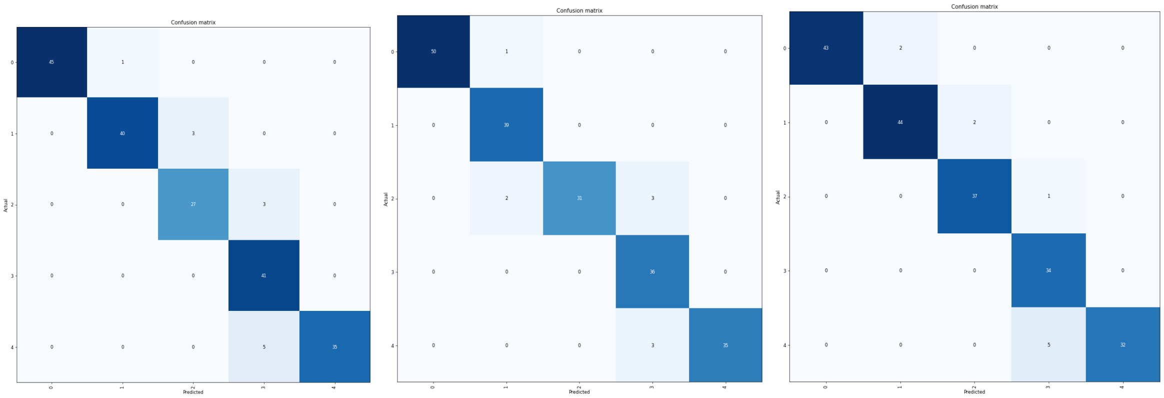
	0	1	2	3	4
0	37	2	0	0	0
1	0	47	0	0	0
2	0	0	43	0	0
3	0	0	2	29	0
4	0	0	0	6	34

	0	1	2	3	4
0	37	2	0	0	0
1	0	47	0	0	0
2	0	0	43	0	0
3	0	0	2	29	0
4	0	0	0	6	34

	0	1	2	3	4
0	41	0	0	0	0
1	0	37	0	0	0
2	0	0	35	0	0
3	0	0	3	43	0
4	0	0	0	5	36

	0	1	2	3	4
0	34	2	1	0	0
1	0	46	0	0	0
2	0	1	32	0	0
3	0	0	3	39	0
4	0	0	0	3	39

Fig 7.Confusion Matrices for different models using Fpz-Cz channel of Sleep-EDF - Sleep Telemetry dataset - (a) ResNet50, (b) MobileNet-v2, (c) AlexNet (d) DenseNet (e)VGG-19 (f) SqueezeNet



	0	1	2	3	4
0	45	1	0	0	0
1	0	40	3	0	0
2	0	0	27	3	0
3	0	0	0	41	0
4	0	0	0	5	35

	0	1	2	3	4
0	50	1	0	0	0
1	0	39	0	0	0
2	0	2	31	3	0
3	0	0	0	36	0
4	0	0	0	3	35

	0	1	2	3	4
0	43	2	0	0	0
1	0	44	2	0	0
2	0	0	37	1	0
3	0	0	0	34	0
4	0	0	0	5	32

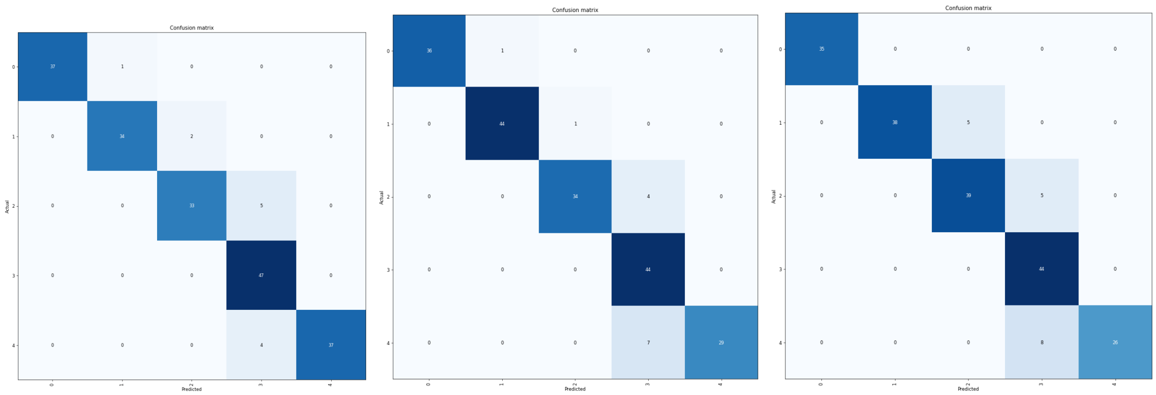


Fig 8.Confusion Matrices for different models using Pz-Oz channel of Sleep-EDF - Sleep Telemetry dataset - (a) ResNet50, (b) MobileNet-v2, (c) AlexNet (d) DenseNet (e)VGG-19 (f) SqueezeNet

Conclusion

The work has implemented and automated the sleep staging pipeline- when given raw, single channel EEG data signals of a patient from the Fpz-Cz of the Sleep-EDF data sets, we classify the five sleep stages (Wake, N1, N2, N3, N4 and REM) using several different Convolutional Neural Network architectures. The preprocessing and sampling implemented helps avoid model bias, ensuring that the models evaluate sleep scoring according to the new AASM standards. All models achieve an accuracy >95% and an f1 score of >0.93 without utilizing any hand-engineered features.<Best Model>

The pipeline utilizes CNNs to extract time-invariant features and learn stage transition rules among sleep stages from EEG epochs. The results also showed that the temporal information learned from the sequence residual learning part helped improve the classification performance. The experiments clearly demonstrated that the models could learn features for sleep stage scoring from different raw single-channel EEGs. The paper also provides an exhaustive overview of the existing literature in the domain, allowing researchers to review the possible approaches of solving the problem. The work also benchmarked the performance of several popular CNN architectures on a consistent sleep stage classification dataset. <> The success of

EEGTransNet can serve as a step on tangible progress not just in the domain of sleep stage classification in particular but also aid the process of integrating AI in medicine.

References

[1] Mohammad Mansour, Fouad Khnaisser, and Hmayag Partamian , “**An Explainable Model for EEG Seizure Detection based on Connectivity Features”** DOI:https://arxiv.org/ftp/arxiv/papers/2009/2009.12566.pdf

[2]. Jean-Marc Fellous, Guillermo Sapiro, Andrew Rossi, Helen Mayberg and Michele Ferrante,
“Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation”
Front. Neurosci., 13 December 2019
DOI:https://doi.org/10.3389/fnins.2019.01346

[3]Amina Adadi and Mohammed Berrada
“ Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”
IEEE Xplore ,September 17, 2018
DOI:https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590

[4] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk and Jocelyn Faubert “**Deep learning-based electroencephalography analysis: a systematic review”**
14 August 2019
Journal of Neural Engineering, Volume 16, Number 5
DOI:https://iopscience.iop.org/article/10.1088/1741-2552/ab260c

[5]A. Supratak, H. Dong, C. Wu and Y. Guo, "**DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG**," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pp. 1998-2008, Nov. 2017, doi: 10.1109/TNSRE.2017.2721116.
DOI:https://ieeexplore.ieee.org/abstract/document/7961240

[6] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, Jimeng Sun **“ SLEEPNET: Automated Sleep Staging System via Deep Learning”**
DOI:https://arxiv.org/pdf/1707.08262.pdf

[7] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews and Yike Guo
“Mixed Neural Network Approach for Temporal Sleep Stage Classification”
DOI:https://arxiv.org/pdf/1610.06421.pdf

[8]Sajad Mousavi, Fatemeh Afghah and U. Rajendra Acharya "**SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach"**
2019 May 7.
DOI: 10.1371/journal.pone.0216456

[9] Faust, Oliver and Razaghi, Hajar and Barika, Ragab and Ciaccio, Edward and Acharya, U Rajendra "**A review of automated sleep stage scoring based on physiological signals for the new millennia"**
April 2019 Computer Methods and Programs in Biomedicine 176
DOI:10.1016/j.cmpb.2019.04.032

[10] Klara Stuburić, Maksym Gaiduk, Ralf Seepold, **"A deep learning approach to detect sleep stages"**
Procedia Computer Science,2020
DOI:**https://doi.org/10.1016/j.procs.2020.09.280**

[11]Huy Phan, Navin Cooray, Oliver Y. Chen, and Maarten De Vos, "**SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging"**
2019 Jan 31.
DOI: 10.1109/TNSRE.2019.2896659

[12] Huy Phan , Oliver Y. Chen, Philipp Koch, Zongqing Lu, Ian McLoughlin, Alfred Mertins, and Maarten De Vos, **"Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning"**
DOI:https://arxiv.org/pdf/1907.13177v3.pdf

[13] Hogeon Seo , Seunghyeok Back , Seongju Lee , Deokhwan Park, Tae Kim, Kyoobin Lee†
"Intra- and Inter-epoch Temporal Context Network (IITNet) Using Sub-epoch Features for Automatic Sleep Scoring on Raw Single-channel EEG"

DOI:https://arxiv.org/pdf/1902.06562v2.pdf

[14] 1. Shepard JW, Buysse DJ, Chesson AL, et al. “**History of the development of sleep medicine in the United States.**”
J Clin Sleep Med. **2005;1(1):61–82.**
DOI:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2413168/

[15]Susan L. Worley, “**The Extraordinary Importance of Sleep: The Detrimental Effects of Inadequate Sleep on Health and Public Safety Drive an Explosion of Sleep Research**”
2018 Dec
DOI:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6281147/

[16] Khald Ali I. Aboalayon ,Miad Faezipour ,Wafaa S. Almuhammadi and Saeid Moslehpour “**Real Time Sleep Detection System Using New Statistical Features of the Single EEG Channel**”
DOI:https://www.researchgate.net/project/Real-Time-Sleep-Detection-System-Using-New-Statistical-Features-of-the-Single-EEG-Channel

[17] **Portable Monitoring in the Diagnosis and Management of Obstructive Sleep Apnea Polysomnography Scoring Manual**
https://sleepdata.org/datasets/homepap/files/m/browser/documentation/HomePAP_PSG_Scoring_Manual.pdf?inline=1

AASM
[18]Iber C, Ancoli-Israel S, Chesson A, Quan SF, editors. 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007. “**The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification.**”
DOI:https://ci.nii.ac.jp/naid/10024500923/

R&K standard

[19] Rechtschaffen A. Brain information service. 1968. “**A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects”**

Visbrain tool

[20] Combrisson, Etienne and Vallat, Raphael and Eichenlaub, Jean-Baptiste and O'Reilly, Christian and Lajnef, Tarek and Guillot, Aymeric and Ruby, Perrine M. and Jerbi, Karim “**Sleep: An Open-Source Python Software for Visualization, Analysis, and Staging of Sleep Data”**
DOI**:**http://journal.frontiersin.org/article/10.3389/fninf.2017.00060

[21]Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun “**Deep Residual Learning for Image Recognition”**
DOI: https://arxiv.org/abs/1512.03385

[22] Hossain, M. S., Amin, S. U., Alsulaiman, M., & Muhammad, G. (2019). “**Applying deep learning for epilepsy seizure detection and brain mapping visualization.”** ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1s), 1-17
DOI: https://arxiv.org/pdf/2007.01276.pdf

[23] H. Phan et al., “Joint classification and prediction CNN framework for automatic sleep stage classification,” IEEE Trans Biomed Eng, vol. 66, no. 5, pp. 1285–1296, 2019.

[24] H. Phan et al., “DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification,” in Proc. EMBC, 2018, pp. 453–456.

[25] H. Phan et al., “Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks,” in Proc. EMBC, 2018, pp. 1452–1455

Sleep edf dataset

[26] B. Kemp et al., “Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG,” IEEE Trans Biomed Eng, vol. 47, no. 9, pp. 1185–1194, 2000.

Sleep edf dataset

[27] A. L. Goldberger et al., “Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, pp. e215–e220, 2000.

Why manual scoring is tiring

[28] R. S. Rosenberg and S. Van Hout, “The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring,” Journal of clinical sleep medicine, vol. 9, no. 01, pp. 81–87, 2013.

physionet

[29] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. E215–e220.

pyEDFlib library

[30] https://github.com/holgern/pyedflib