Published in final edited form as:
Clin Neurophysiol. 2019 January ; 130(1): 77–84. doi:10.1016/j.clinph.2018.10.012.

Detecting Abnormal Electroencephalograms Using Deep Convolutional Networks

K.G. van Leeuwen[#1,2], **H. Sun, PhD**[#1], **M. Tabaeizadeh, MD**[1], **A.F. Struck, MD**[3], **M.J.A.M van Putten, MD, PhD**[2,4], and **M.B. Westover, MD, PhD**[1]

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA [2]University of Twente, Enschede, the Netherlands [3]Department of Neurology, Wisconsin Hospital and Clinics, Madison, WI, USA [4]Department of Neurology and Clinical Neurophysiology, Medisch Spectrum Twente, Enschede, the Netherlands

[#] These authors contributed equally to this work.

Abstract

Objectives: Electroencephalography (EEG) is a central part of the medical evaluation for patients with neurological disorders. Training an algorithm to label the EEG normal vs abnormal seems challenging, because of EEG heterogeneity and dependence of contextual factors, including age and sleep stage. Our objectives were to validate prior work on an independent data set suggesting that deep learning methods can discriminate between normal vs abnormal EEGs, to understand whether age and sleep stage information can improve discrimination, and to understand what factors lead to errors.

Methods: We train a deep convolutional neural network on a heterogeneous set of 8,522 routine EEGs from the Massachusetts General Hospital. We explore several strategies for optimizing model performance, including accounting for age and sleep stage.

Results: The area under the receiver operating characteristic curve (AUC) on an independent test set (n = 851) is 0.917 marginally improved by including age (AUC=0.924), and both age and sleep stages (AUC= 0.925), though not statistically significant.

Conclusions: The model architecture generalizes well to an independent dataset. Adding age and sleep stage to the model does not significantly improve performance.

Significance: Insights learned from misclassified examples, and minimal improvement by adding sleep stage and age suggest fruitful directions for further research.

Address for correspondence: M. Brandon Westover, MD, PhD, Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA, Phone: +1 650-862-1154, mwestover@mgh.harvard.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Electroencephalograms (EEG); convolutional neural networks (CNN); computer aided diagnosis (CAD); deep learning; epilepsy; clinical neurophysiology

1. Introduction

Electroencephalography (EEG) can be used to detect the abnormal patterns of brain electrical activity present in a broad range of neurological and medication conditions. For example, EEGs of patients with epilepsy often exhibit characteristic “epileptiform” discharges (epileptic spikes or sharp-waves) (Schomer and Da Silva 2012). Lesions, such as strokes or hemorrhages, can result in asymmetry across left and right hemispheres (Agius Anastasi et al. 2017; Jordan 2004; van Putten 2007). Patients with depressed levels of consciousness exhibit generalized slowing of EEG rhythms or burst suppression patterns (Young 2000; Kaplan 2004; Schomer and Da Silva 2012) Metabolic encephalopathy from acute liver failure can cause abnormalities such as triphasic waves (Boulanger et al. 2006; Foreman et al. 2016).

Routine EEGs, brief recordings lasting typically 20–30 minutes, play an important part in diagnosing these conditions. However, there are several important difficulties in determining whether an EEG is normal or abnormal. First, inter-rater agreement is moderate. For example, a spike-and-wave discharge can be brief and small in amplitude, barely distinguishable from the background. Asymmetries of the EEG background similarly range from obvious to subtle. Such ambiguities lead to imperfect agreement among clinical neurophysiologists and inconsistent interpretations of the same EEG. For example, six board-certificated neurophysiologists classified 300 EEGs from a general clinically heterogeneous population (≥ 1-year-old) as normal vs containing seizures or epileptiform discharges and achieved an agreement (Fleiss’s kappa) of 55% (Grant et al. 2014). For neonates with hypoxic ischemic encephalopathy, three pediatric neurophysiologists reviewed 60 EEGs and categorized them as normal vs. abnormal (Wusthoff et al. 2017) and achieved an agreement (Fleiss’s kappa) of 49%. Second, determining whether an EEG is normal or abnormal is also time-consuming. Many abnormal patterns are intermittent, thus the interpreting neurologist must review the entire EEG to perform an adequate EEG analysis.

Deep neural networks, including convolutional neural networks (CNN) have recently been used for EEG classification tasks (See Supplementary Appendix A, Table A1, for a review). Schirrmeister et al. recently described a convolutional neural network (CNN) that classifies EEGs as normal vs abnormal, using 20 minutes of 21-channel EEG from 3017 subjects in the TUH dataset (Obeid and Picone 2016). The network was trained in an end-to-end manner without hand-crafted features. On a test set of 277 routine EEG recordings, the network achieved an accuracy of 84.8%. An analysis of the misclassified cases suggested that future work might be able to improve performance by considering patient age and state (awake vs asleep).

In this study, we develop a convolutional neural network to classify normal and abnormal EEG based on 8,522 routine EEGs. Our model builds on the prior work of Schirrmeister et al. Beyond minor technical enhancements, our efforts extend this prior work in four important ways. First, by training on a larger and more heterogeneous dataset, our work validates the previous results, and provides a more stringent test of the extent to which detecting generic EEG abnormalities is learnable by CNN models. This addresses the "replication crisis" in science (Schooler 2014), which is particularly important in the current climate of deep neural networks, where models do not always show similar results when used on a different dataset. Second, we systematically explore whether including age and sleep stages into the model improves performance. In clinical practice, age contributes to EEG interpretation. For example, normal elderly patients show more slowing of the posterior alpha rhythm and decreased overall amplitude relative to younger patients (Mander et al. 2017). As another example, for children under 1-year-old, the posterior dominant rhythm is typically slower, in the 3 – 4 Hz range, than the typical adult range of 8–12Hz and vertex waves seen during sleep tend to be very high in amplitude and sharply peaked ("spiky") relative to vertex waves in adults (Grigg-Damberger et al. 2007; Ebersole and Pedley 2003). In addition, patterns that would be abnormal in an awake EEG may be normal while asleep. For example, generalized slowing, a sign of encephalopathy in an awake patient, is normal during drowsiness and sleep. Third, we explore various ways of pooling temporal information to come to a final overall decision about whether an EEG is normal or abnormal. Finally, we analyze model prediction errors. Our analysis suggests future directions that may be fruitful for improving CNN models, and provides information regarding the degree of uncertainty in the training labels, which govern the ultimate performance ceiling of supervised machine learning approaches for this classification task.

2. Methods

2.1 Dataset

The Partners Institutional Review Board approved retrospective analysis of the dataset without requiring additional consent for its use in this study. A database of 8,522 routine EEGs from the Department of Neurology in Massachusetts General Hospital was collected from 2012 to 2016. All EEG recordings were recorded using the standard international 10–20 EEG system. All EEGs included had a minimum duration of 20 minutes (see Figure 1). Each EEG was reviewed by a minimum of two experienced clinical electroencephalographers, who described their findings in semi-structured EEG reports that were filed in the electronic medical record. All reports include a header which declares the EEG as being either "normal" or "abnormal". The label "normal" or "abnormal" is extracted from each report and used as target labels to be predicted for our study. Only subjects between ages 18 and 85 years old are included in the present study. Subject characteristics are summarized in Table 1.

2.2 Preprocessing

EEGs are resampled to 100 Hz and clipped from −800 to 800 mV to reject unphysiologically extreme values. A longitudinal bipolar ("double banana") montage is constructed from the original reference recording montage. Each 6 seconds of EEG is automatically assigned a label of 'good quality' EEG, 'flatline', or 'extreme values'. Using these labels, a segment of 15 minutes consisting of at least 90% 'good quality' data is extracted.

The subjects are randomly split into a train (7671 EEG recordings) and test set (851 EEG recordings) with likelihood of 0.9 and 0.1 respectively, in which no subject overlap is present. A sample of one tenth of the train set is taken as the validation set (767 EEG recordings).

2.3 Network architecture

The architecture of our CNN is adapted from the network architecture of Schirrmeister et al. (Schirrmeister et al. 2017b; Schirrmeister et al. 2017a) (Figure 2). For each 1-minute EEG segment (6000 × 18), we obtain NxM values, with N=5400 the number of time points and M=200 the number of filters. We include an extra average pooling layer to take the average along the time point axis before the last classification layer. This layer enables us to add new features to the learned EEG features before the linear classification layer.

The network is trained using the Adam optimizer (Kingma and Ba 2014). In each minibatch of 64 samples (1 minute of EEG), the samples are augmented by flipping EEG channels of the right and left hemispheres with probability 50% to prevent overfitting to one hemisphere. The training is further regularized with batch normalization, dropout and early-stopping.

2.4 Including Age

Subject age is incorporated into the model in two different ways. First the normalized age is added to the 200-dimensional feature vector right after the average pooling layer, making it a feature vector of 201 dimensions, as seen in Figure 3A. Secondly, we perform transfer learning (Sharif et al. 2014), in which the model trained on all ages is fine-tuned on subgroups with about a ten-year span (18–29, 30–39, 40–49 years). We experiment with the number of layers in which to allow for parameter updates.

2.5 Including Sleep Stage

To assess the impact of providing information about sleep stages, we use a recently published algorithm that performs at a level similar to human experts in assigning sleep stages to consecutive 30 second epochs of EEG (Sun et al. 2017). The model outputs a probability for stages NREM1, NREM2, NREM3, REM or awake. The probabilities over the EEG segment are averaged over 1 minute to match the input size of the current network. This vector of five probability values is concatenated with the 200-dimensional feature vector and age, producing a 206-dimensional input for the final classification layer (Figure 3B).

2.6 Aggregation

The model generates a probability value for each minute of EEG, therefore we get multiple predicted probabilities per 15-minute EEG sample. These multiple predictions of each 15-minute segment are combined to obtain a final single label of “normal” or “abnormal”, which can be compared with the overall impression given to the EEG in the clinical EEG report. Our baseline approach is to average these multiple predicted probabilities from each minute of EEG into a single abnormality probability for each 15-minute EEG sample.

As an alternative to the baseline approach, we experiment with a long-short-term-memory layer (LSTM) to consider the temporal domain in the final classification. The inputs of the LSTM are the feature vectors of length 200 from the convolutional layers before the linear classifier. Instead of having an average pooling layer of 1 × 5400, we use an average pooling of 1 × 600, by which we obtain a feature vector for each 6 seconds of input. We perform hyperparameter tuning by changing the number of hidden layers (1 or 2), the hidden layer size (32 or 64), drop-out rate (0 or 0.25), and the direction (unidirectional or bidirectional). The LSTM is trained separately from the main convolutional model.

2.7 Statistical Analysis and Model Evaluation

To compare and evaluate the performance of the trained models, we use area under the receiver operating characteristic (AUC) for the train, validation and test sets. Statistical significance of differences between AUCs of different models is tested via the permutation test with 5000 repetitions and significance level of 0.05. Accuracy and specificity at a level of 90% sensitivity are also calculated.

To gain insight into the reasons for EEGs that are misclassified by the final model, we use five approaches. First, we create a 2-dimensional visualization using t-SNE of the output after the average pooling layer. This gives us insights in which EEGs are misclassified. Secondly, weights of the final linear layer are plotted to visualize the importance of age and sleep stage features. Third, the accuracy per age group is reviewed comparing the baseline model with the model including age and sleep stage. Fourth, we examine the clinical reports to identify factors that might differentiate misclassified from correctly classified EEGs. For this analysis EEG reports are analyzed for word frequencies. The frequencies of each unique word in the entire corpus of EEG reports is determined and compared in misclassified vs correctly classified samples. The ratio of word frequencies in misclassified and in correctly classified are calculated, with ratios > 1 indicating words that appear more frequently in misclassified samples, and <1 indicating words more frequent in correctly classified samples. Student t-tests are performed to assess statistical significance of deviations of the ratio from 1. Fifth, a random subset of misclassified EEG samples is reviewed manually to generate hypotheses about the reasons for misclassification.

3. Results

3.1 Generalization of algorithm

Our baseline model and the original Schirrmeister et al model perform similarly (AUC 0.914 vs 0.917, p = 0.45) on the MGH dataset, as shown in Table 2. In Supplementary Appendix B, Figure B1, it can be seen that train, validation and test set have a similar loss indicating the network is not overfitting to the train set.

3.2 Including Age

Adding age to the feature vector after the average pooling layer yields a small but statistically insignificant performance improvement compared with the baseline model without age included (Table 3).

We also attempt to improve performance by creating distinct models for different age groups, by using transfer learning to adapt the baseline model. Table 4 gives the results for three age groups (18–29, 30–39 and 40–49 years old) used to separately fine tune the baseline model that is pre-trained on all training data (18–85 years old). The results are compared to the baseline model tested on each age group separately. As shown in Table 4, accounting for age in this way does not yield any statistically significant performance gains.

3.3 Including sleep stages

Most routine EEGs contain periods of wake, drowsiness (stage N1 sleep), and stage N2 sleep. When the model is trained including age and sleep stage information concatenated to the feature vector feeding into the linear classification layer, the results are again statistically indistinguishable from the baseline model (p-value of 0.35), as seen in Table 3. Figure 4 visualizes the performance differences of the baseline model with and without age and sleep stage accounted for.

Figure 5 shows the weights of the final classifying linear layer, of which 200 are descriptive of the EEG, five for sleep stage probabilities and one for age. Weights close to zero have little influence on the final probability value for the EEG being abnormal and a positive value promotes abnormality. The figure shows that the weight learned by the final layer for age is only slightly positive, favoring abnormality when the age is larger. The sleep stages all have larger (negative or positive) weights than the mean EEG weights, except for the awake feature.

3.4 Aggregation

Training an LSTM to integrate temporal information across the duration of the EEG does not produce any significant performance gain over the baseline method of simple averaging. The LSTM providing the best results (2 hidden layers, dropout 0.25, hidden units 64, bidirectional), leads to an AUC the same as the baseline as can be seen in Table 3.

3.5 Evaluation of misclassification: feature embedding map

The t-SNE plot (Figure 6) shows how features of normal and abnormal EEGs overlap in the projected feature space. The threshold used to distinguish normal and abnormal is 0.5, which gives 91.7% specificity and 73.9% sensitivity. The confusion matrix is shown in Table 5. There are only 32 normal EEG samples (3.8% of all test samples) that have been misclassified as abnormal (upper right, false positive). The opposite is more prevalent (bottom left, false negative) and can be understood by the fact that these points appear in a region of feature space that is near many normal samples.

3.6 Evaluation of misclassification: age related error

Figure 7 demonstrates that the misclassification percentage over age only minimally changes when including age in the model. The largest performance gain is seen in the age group from 18 to 29 years old, where the accuracy increases by 3%, from 81% to 84% when including age. Between age groups the maximum accuracy deviation is about 10%, with the highest accuracy of 87.2% seen in the age group of 70 to 85.

3.7 Evaluation of misclassification: EEG report word frequency analysis

The word analysis (Table 6) of the EEG text reports for the test set shows which words are more frequent in the misclassified cases over the correctly classified cases. If the ratio is greater than 1 the word is more prevalent in the misclassified sample, indicating these cases are harder to classify. Abnormal cases with words related to sleep are more often misclassified, similar to findings of Schirrmeister et al (Schirrmeister et al. 2017a). Cases with prominent nonphysiological signal artifacts are more likely to be misclassified as abnormal. In misclassified cases, the scoring neurologist uses expressions of doubt or uncertainty 1.93 and 2.35 times more often than in correctly classified samples of abnormal and normal samples respectively. EEG reports including words related to slowing or spike wave discharges, but labeled as normal, were more often predicted to be abnormal and were thus misclassified, presumably because such EEGs contained "suspicious" but not sufficiently distinct or obvious to be considered definitively abnormal. In Supplementary Appendix C, Table C1, this word analysis is compared to the word analysis of the baseline model showing minimal deviations.

4. Discussion

In this study, we have developed a deep convolutional neural network-based that detects abnormal EEGs, using a large and heterogeneous clinical routine EEG dataset containing 8,522 EEGs (MGH data). The baseline model performs remarkably well, despite being tested on a wide range of ages (18–85), and despite being provided none of the explicit contextual information that is typically utilized by clinical experts such as patient age and sleep stage. Somewhat surprisingly, we find that several different ways of attempting to include information about age and sleep stage are unable to further improve classification performance with statistical significance, nor does using a recurrent neural network (LSTM) to combine information across time improve performance over simple averaging. Our work builds on that of Schirrmeister et al. (Schirrmeister et al. 2017b; Schirrmeister et al. 2017a) Performance of the two models are similar, and demonstrate the generalizability and robustness of the CNN-based approach on independent large datasets.

4.1 Age

We researched different ways of incorporating the subject's age in the predictive model, but these all failed to enhance performance. This may be because age was only important in distinguishing normal from abnormal for a small number of EEGs, i.e. that neurologists would have also labeled the EEGs similarly even without knowing the age. Alternatively, the model may have implicitly learned to take age-related features into account in reaching a determination of normal vs abnormal. Figure 7 shows the largest accuracy gain when including age is found in the youngest age group (18–29 years old). As younger brains are still more in development, including age in the analysis might be more beneficial for this subgroup. Because of this, incorporating age in the model when including children's EEGs might be more successful as children EEGs deviate more from each other due to age.

4.2 Sleep stage

For the first time, this study pursues to include sleep stage information in the classification model. Word frequency analysis suggests that sleep is an important factor underlying misclassification, though our attempts to include explicitly sleep stage information in the model did not improve performance. The word analysis as in Supplementary Appendix C, shows that the appearance for sleep-related words in misclassified samples barely change when adding sleep stage to the model indicating its minimal impact. ‘Covariate drift’ is one possible explanation, as the algorithm used for determining sleep stages was developed using overnight sleep EEGs from a sleep lab rather than the routine EEGs in this study. In support of this, we found that some misclassified EEG segments had been labeled by the sleep staging algorithm as REM when in fact the patient was awake with eyes open. Similarly, the algorithm occasionally mistakes periods of hyperventilation (which tend to induce slow oscillations) or abnormal background slowing with N3 sleep.

4.3 Aggregation

The initial approach to combine predictions of 1-minute pieces of EEG into a final label for the 15-minute EEG segment was to average probability values. We trained an LSTM to contemplate the time domain and the intermittent characteristics of many abnormalities, however, without significant improvement. The LSTM was trained separately from the convolutional network. It might improve results if it were to be included to become an end-to-end trained model.

4.4 Misclassification analysis

We compared word frequencies in EEG text reports to understand which features recognized by expert EEG readers might underlie classification errors and successes. By manually evaluating confidently correctly detected abnormal EEGs, we find that theta and delta slowing are prominent common abnormalities, congruent with the word frequency analysis. Word frequency analysis also suggests that artifacts bias raters toward the abnormal class. We also note that muscle artifacts are a source of occasional errors; the model’s estimated probability of the EEG being abnormal often rises during periods of muscle artifact. These findings suggest that artifact reduction methods might be beneficial to increase specificity. One of the most outstanding word clusters was the one expressing doubt by EEG reviewers. This raises the possibility of further improving discrimination by considering label noise. When phrases describing abnormalities (slowing, spike, discharge, etc.) were used in the reports of normal EEGs, EEGs were more often misclassified, giving further support for label noise being an issue. Interrater agreement in EEG reading is well known to be imperfect, and could thus explain some of the error (van Donselaar et al. 1992; Azuma et al. 2003).

5. Conclusion

We developed a deep convolutional neural network to classify routine EEG recordings as normal or abnormal, using 8,522 EEGs from the MGH dataset. On the MGH dataset we achieved AUC at 0.917. Including age improved the AUC to 0.924, and including both age and sleep stages to 0.925, though not significant. The analysis of factors underlying classification errors suggests promising directions for further improving model performance and interpretability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Disclosures and Conflict of interest:

MBW and HS receive funding from NIH-NINDS (1K23NS090900, 1R01NS102190–01). KVL reports no disclosures. MVP is a co-founder of Clinical Science Systems, a manufacturer of EEG analysis software. AS reports no disclosures.

References:

Agius Anastasi A, Falzon O, Camilleri K, Vella M, & Muscat R Brain symmetry index in healthy and stroke patients for assessment and prognosis. Stroke Res Treat 2017 10.1155/2017/8276136

Azuma H, Hori S, Nakanishi M, Fujimoto S, Ichikawa N, & Furukawa TA An intervention to improve the interrater reliability of clinical EEG interpretations. Psychiatry Clin Neurosci 2003, 57(5), pp. 485–489. 10.1046/j.1440-1819.2003.01152.x. [PubMed: 12950702]

Bashivan P, Rish I, Yeasin M, & Codella N Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks; 2015 arXiv:1511.06448.

Biswal S, Kulas J, Sun H, Goparaju B, Westover MB, Bianchi MT, et al. SLEEPNET: Automated Sleep Staging System via Deep Learning; 2017 arXiv:1707.08262.

Boulanger JM, Deacon C, Lécuyer D, Gosselin S, & Reiher J Boulanger. Triphasic waves versus nonconvulsive status epilepticus: EEG distinction. Can J Neurol Sci 2006, 33(2), pp.175–180. 10.1017/S0317167100004935 [PubMed: 16736726]

Bozal A Personalized Image Classification from EEG Signals using Deep Learning. Thesis, Universitat Politecnica de Catalunya, 2017.

van Donselaar CA, Schimsheimer RJ, Geerts AT, & Declerck AC Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. Arch Neurol, 1992, 49(3), pp.231–237. 10.1001/archneur.1992.00530270045017. [PubMed: 1536624]

Ebersole JS & Pedley TA Current practice of clinical electroencephalography. 3rd ed. Lippincott Williams & Wilkins; 2003 10.1046/j.1468-1331.2003.00643.x

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, & Thrun S Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542(7639), p.115 10.1038/nature21056 [PubMed: 28117445]

Foreman B, Mahulikar A, Tadi P, Claassen J, Szaflarski J, Halford JJ et al., 2016 Generalized periodic discharges and “triphasic waves”: A blinded evaluation of inter-rater agreement and clinical significance. Clin Neurophysiol 2016, 127(2), pp.1073–1080. 10.1016/j.clinph.2015.07.018 [PubMed: 26294138]

Grant AC, Abdel-Baki SG, Weedon J, Arnedo V, Chari G, Koziorynska E, et al., EEG interpretation reliability and interpreter confidence: a large single-center study. Epilepsy Behav 2014, 32, pp. 102–107. 10.1016/j.yebeh.2014.01.011 [PubMed: 24531133]

Grigg-Damberger M, Gozal D, Marcus CL, Quan SF, Rosen CL, Chervin RD et al. The visual scoring of sleep and arousal in infants and children. J Clin Sleep Med 2007, 3(2), pp.201–240. [PubMed: 17557427]

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016, 316(22), pp.2402–2410. 10.1001/jama.2016.17216. [PubMed: 27898976]

Halford JJ, Arain A, Kalamangalam GP, LaRoche SM, Leonardo B, Basha M et al. Characteristics of EEG interpreters associated with higher interrater agreement. J Clin Neurophysiol 2017, 34(2), pp. 168–173. https://dx.doi.org/10.1097%2FWNP.0000000000000344 [PubMed: 27662336]

Halford JJ, Schalkoff RJ, Zhou J, Benbadis SR, Tatum WO, Turner RP et al. Standardized database development for EEG epileptiform transient detection: EEGnet scoring system and machine learning analysis. J Neurosci Methods 2013, 212(2), pp.308–316. 10.1016/j.jneumeth.2012.11.005 [PubMed: 23174094]

Hosseini MP, Soltanian-Zadeh H, Elisevich K, & Pompili D Cloud-based Deep Learning of Big EEG Data for Epileptic Seizure Prediction; 2017 arXiv:1702.05192.

Jirayucharoensak S, Setha P-N & Israsena P EEG-Based Emotion Recognition Using Deep Learning Network with Principal Component Based Covariate Shift Adaptation. Sci World J, 2014, 10.1155/2014/627892

Jordan KG Emergency EEG and Continuous EEG Monitoring in Acute Ischemic Stroke. J Clin Neurophysiol 2004, 21(5), pp.341–352. doi: 10.1097/01.WNP.0000145005.59766.D2 [PubMed: 15592008]

Kaplan PW The EEG in metabolic encephalopathy and coma. J Clin Neurophysiol 2004, 21(5), pp. 307–318. [PubMed: 15592005]

Kingma DP & Ba J, 2014 Adam: A method for stochastic optimization; 2014 arXiv:1412.6980.

Mander BA, Winer JR & Walker MP Review Sleep and Human Aging. Neuron 2017, 94(1), pp.19–36. 10.1016/j.neuron.2017.02.004. [PubMed: 28384471]

Ni Zhaoheng, Yuksel Ahmet Cem, Ni Xiuyan, Mandel Michael I., and L.X., 2017 Confused or not Confused?: Disentangling Brain Activity from EEG Data Using Bidirectional LSTM Recurrent Neural Networks. In: Proc: 8th ACM Int Conf on Bioinform, Comput Biol, and Health Inform, pp. 241–246. ACM, 2017 10.1145/3107411.3107513

Obeid I, Picone J The Temple University Hospital EEG Data Corpus. Front. Neurosci 2016,10,196 [dataset] doi: 10.3389/fnins.2016.00196 [PubMed: 27242402]

Schirrmeister R, Gemein L, Eggensperger K, Hutter F, & Ball T Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. IEEE Signal Process in Med and Biol Symp, Philadelphia, PA, 2017a, pp. 1–7. doi: 10.1109/SPMB.2017.8257015

Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glasstetter M, Eggensperger K, Tangermann M 2017b Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 2017b, 38(11), 5391–5420. 10.1002/hbm.23730 [PubMed: 28782865]

Schomer DL & Da Silva FL, 2012 Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins.

Schooler JW Metascience could rescue the 'replication crisis'. Nature News 2014, 515(7525), 9. doi: 10.1038/515009a

Sharif Razavian A, Azizpour H, Sullivan J, & Carlsson S CNN Features off-the-shelf: an Astounding Baseline for Recognition. In: Proc: IEEE Conf on Comp Vis Pattern Recognit Workshops (CVPR), 2014, pp. 806–813.

Stober S, Sternin A, Owen AM, & Grahn JA Deep Feature Learning for EEG Recordings; 2015 arXiv: 1511.04306

Stober S, Cameron DJ & Grahn JA Classifying EEG recordings of rhythm perception. 15th Int Soc Music Inf Retr Conf, (Ismir), 2014, pp.649–654.

Sun H, Jia J, Goparaju B, Huang GB, Sourina O, Bianchi MT, & Westover MB Large-Scale Automated Sleep Staging. Sleep 2017, 40(10). 10.1093/sleep/zsx139

Thodoroff P, Pineau J & Lim A Learning Robust Features using Deep Learning for Automatic Seizure Detection In: Proc: Mach Learn Healthc Conf (MLHC) 2016, JMLR W&C Track, vol 56, pp.178–190, Los Angeles https://arxiv.org/abs/1608.00220

Turner JT, Page A, Mohsenin T, & Oates T Deep Belief Networks used on High Resolution Multichannel Electroencephalography Data for Seizure Detection. In: Proc: AAAI Spring Symp Ser, 2014, pp.75–81. http://www.aaai.org/ocs/index.php/SSS/SSS14/paper/viewPDFInterstitial/7747/7787.

van Putten MJ The revised brain symmetry index. Clin Neurophysiol 2007, 118(11), pp.2362–2367. 10.1016/j.clinph.2007.07.019 [PubMed: 17888719] van Putten MJAM, Hofmeijer J, Ruijter BJ, Tjepkema-Cloostermans MC Deep Learning for outcome prediction of postanoxic coma In: Proc: Eskola H, Väisänen O, Viik J, Hyttinen J (eds) EMBEC & NBC 2017 EMBEC 2017, NBC 2017. IFMBE Proceedings, vol 65 Springer, Singapore 10.1007/978-981-10-5122-7_127

Wulsin DF, Gupta JR, Mani R, Blanco JA, & Litt B Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. J Neural Eng 2011, 8(3), 036015 10.1088/1741-2560/8/3/036015 [PubMed: 21525569]

Wusthoff CJ, Sullivan J, Glass HC, Shellhaas RA, Abend NS, Chang T et al., Interrater agreement in the interpretation of neonatal electroencephalography in hypoxic ischemic encephalopathy. Epilepsia 2017, 58(3), pp.429–435. 10.1111/epi.13661 [PubMed: 28166364]

Young GB The EEG in coma. J Clin Neurophysiol 2000, 17(5), pp.473–485. [PubMed: 11085551]

Highlights

1. We validate a convolutional neural network identifying abnormal EEGs in a large diverse set of 8522 EEGs.
2. Including age and sleep stage in the model results in minimal performance gain.
3. Extensive prediction error analysis reveals promising future research directions.

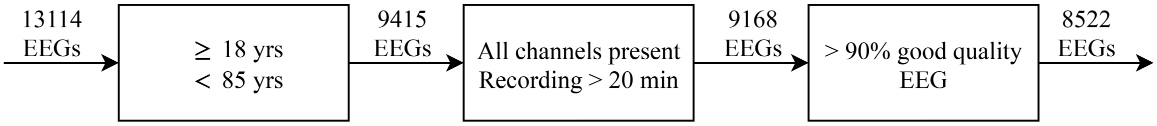

Figure 1:

Flowchart of the data selection process.

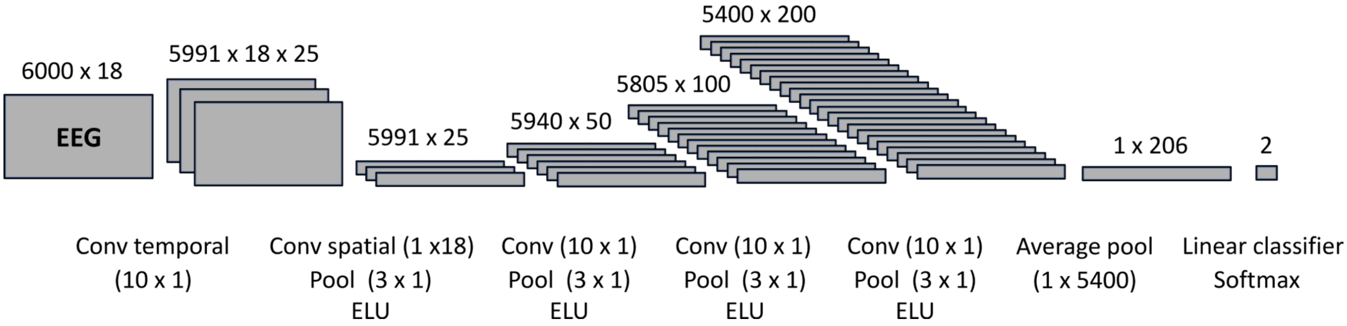
6000 x 18
EEG
5991 x 18 x 25
5991 x 25
5940 x 50
5805 x 100
5400 x 200
1 x 206
2
Conv temporal (10 x 1)
Conv spatial (1 x18) Pool (3 x 1) ELU
Conv (10 x 1) Pool (3 x 1) ELU
Conv (10 x 1) Pool (3 x 1) ELU
Conv (10 x 1) Pool (3 x 1) ELU
Average pool (1 x 5400)
Linear classifier Softmax

Figure 2:

Baseline model. Convolutional neural network starting with separate temporal layer and spatial convolutional layers, followed by a pooling layer. This is followed by three similar convolution+pooling blocks including dropout and batch normalization. An extra average pooling layer was added before the last fully connected linear layer with a softmax activation. The numbers between brackets represent the kernel sizes.

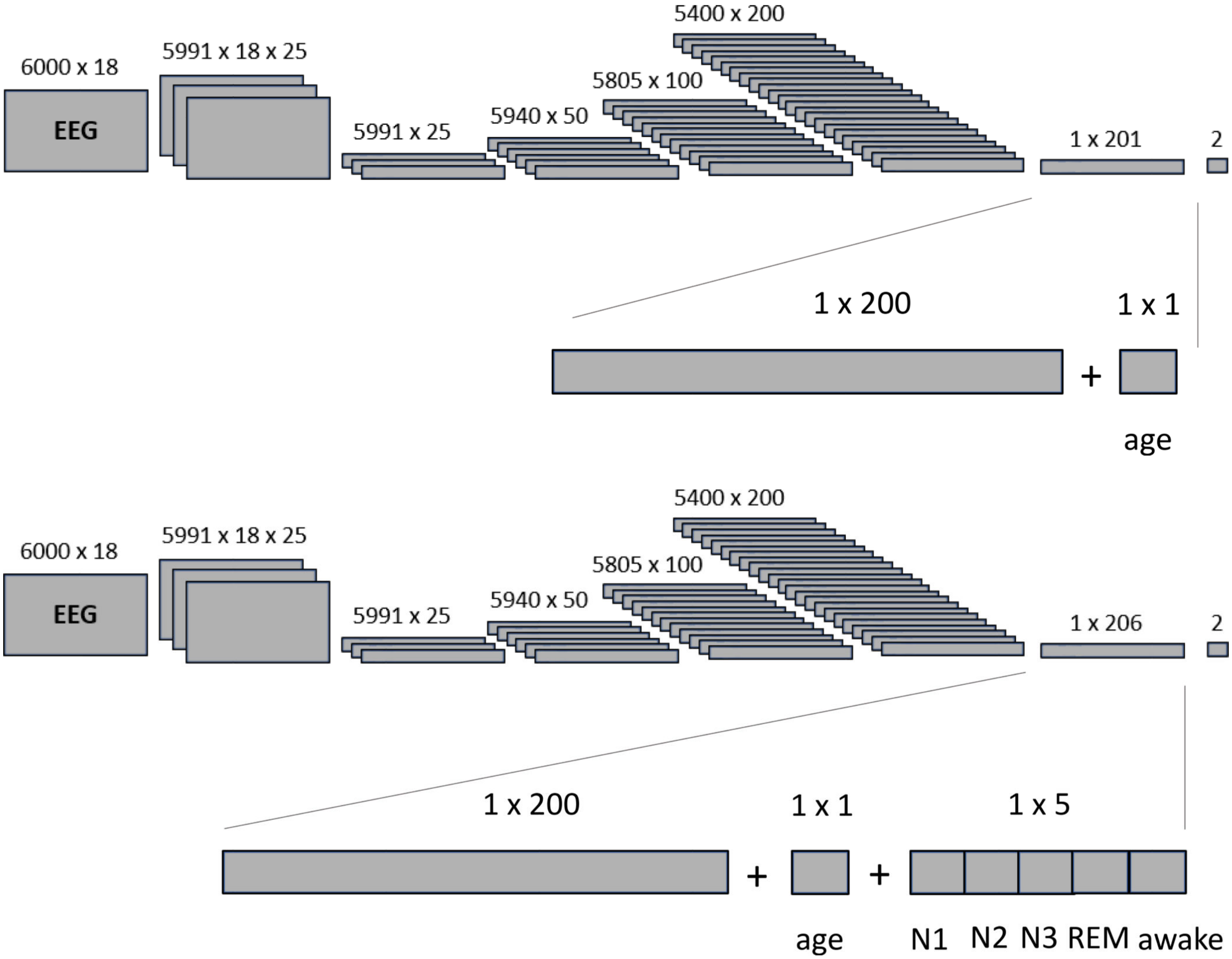

Figure 3:

Network architecture including age (a) and sleep stages (b). After the average pooling layer, the age (normalized age) and sleep stage features (probability value for each sleep stage) were concatenated with the EEG feature vector (1 × 200). The combined vector was the input for the classification layer.

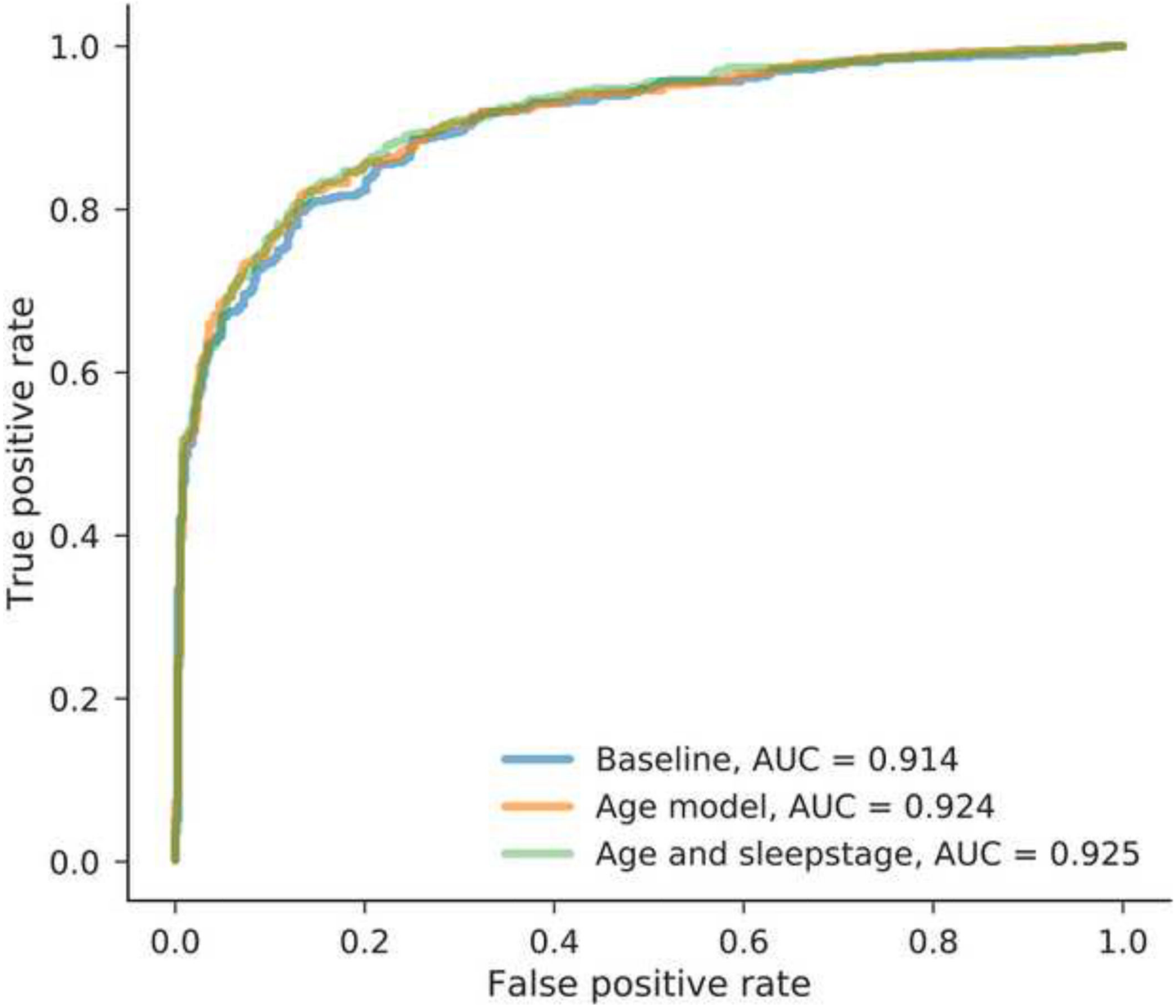

Figure 4:

ROCs on the test set for the baseline model, the model including age, and the model including both age and sleep stages. ROCs are similar, and AUC differences are not statistically significant.

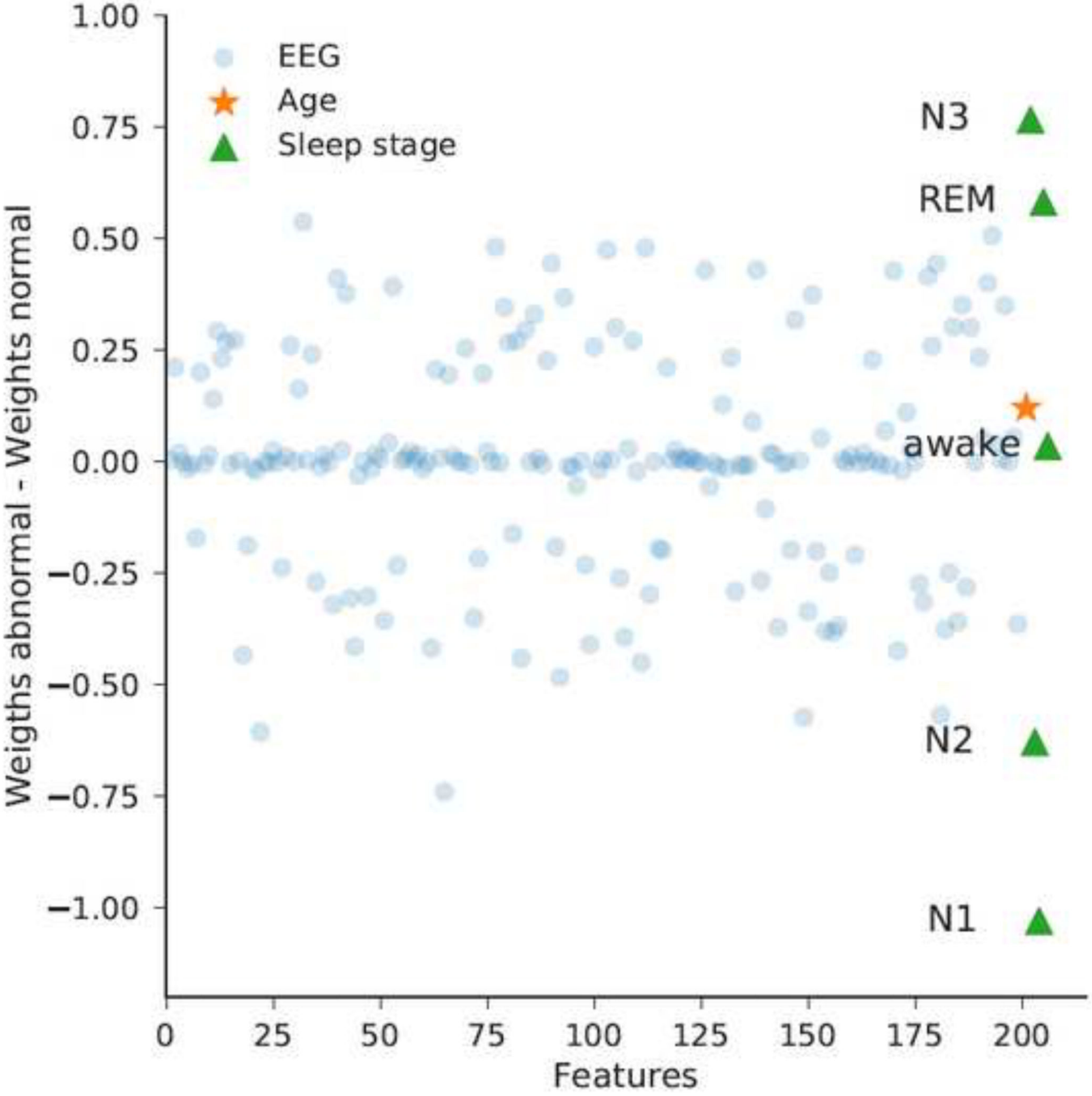

Figure 5:

Weights of the linear classification layer. Weights for the normal class were deducted from the weights for the abnormal class. Features with weights close to zero have little influence on the final prediction of the model. A feature with a high positive weight promotes abnormality.

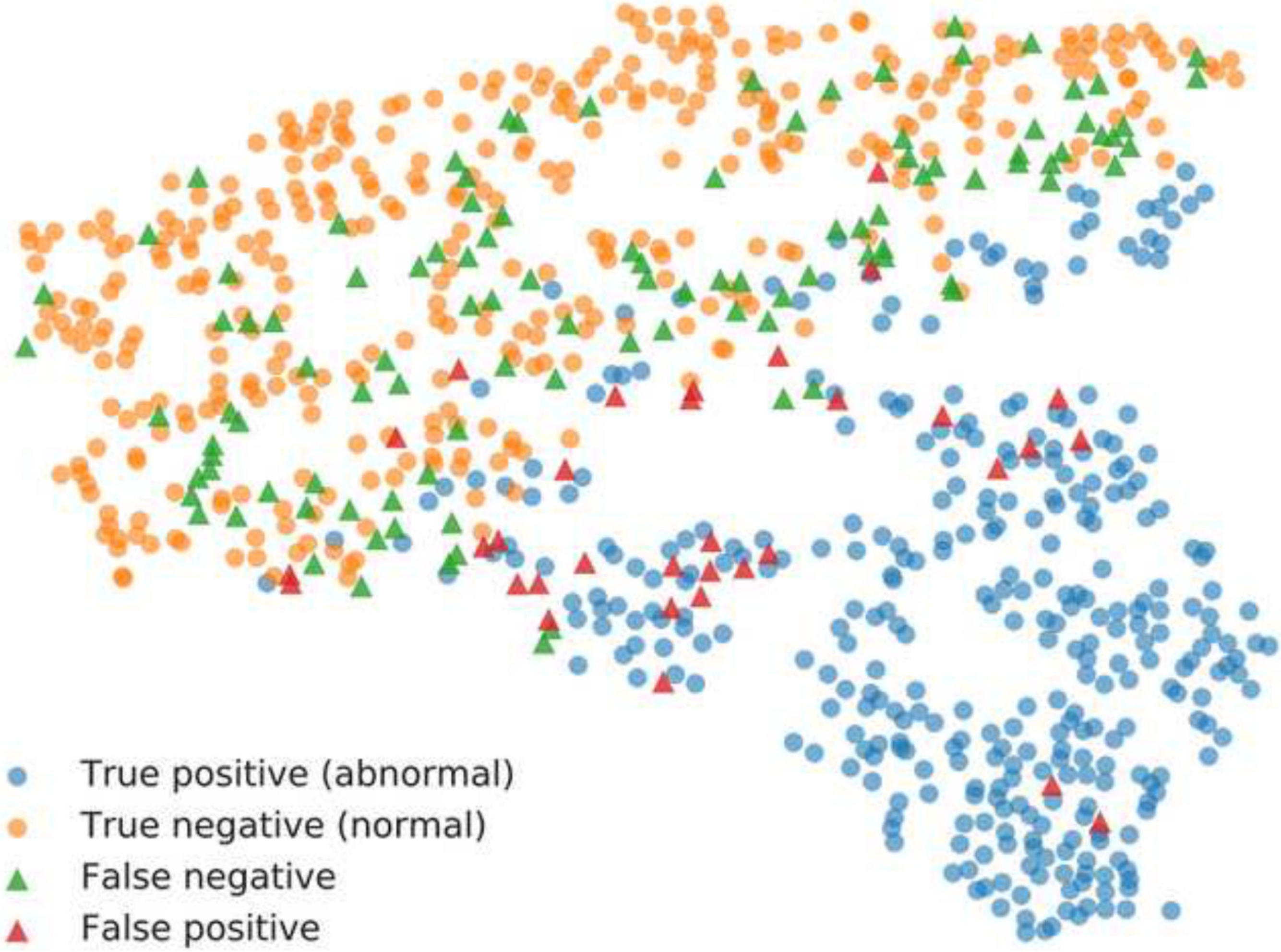

Figure 6:

T-SNE plot of the feature vectors before classification of the baseline model applied to the test set. Dots show correctly classified EEGs and triangles misclassified EEGs. The abnormal class is considered the positive class. Classification threshold of 0.5.

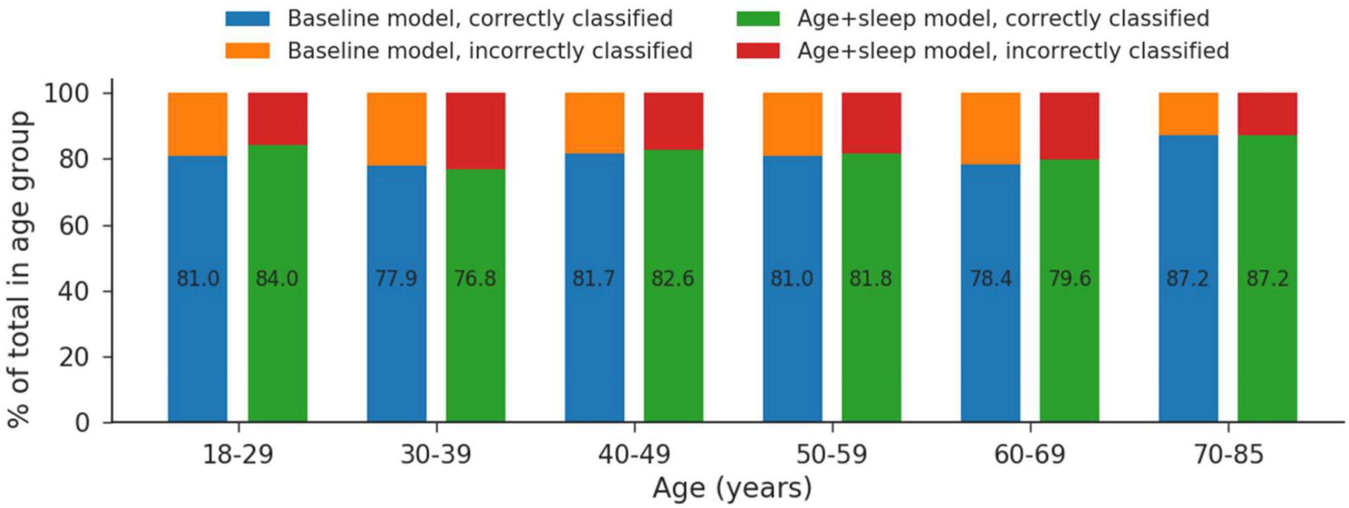

Figure 7:

Bar plot showing the percentages of correctly and incorrectly classified samples (accuracy) per age group of the baseline model and the model including age and sleep stage. Accuracy gain when adding age and sleep stage is highest in the 18–29 years group. Overall accuracy is highest in the 70–85 years group.

Table 1

Subject Characteristics in the MGH Dataset

Variable	Train	Test
Number of EEGs	7671	851
Number of subjects	6465	835
Sex, male (percentage)	3875 (50.5%)	425 (50.0%)
Mean age ± std (year)	52.1 ± 19.2	51.5 ± 19.2
Abnormal/normal ratio	0.56	0.53

Table 2

Results of test set of the MGH dataset, with predictions made by the two baseline models

		Specificity at 90% Sensitivity (%)	Accuracy (%)	AUC	p-value AUC
a	Our baseline model	74.8	81.6	0.914	0.45
b	Schirrmeister’s model	74.1	81.6	0.917	

Table 3

Results on the MGH test set with and without age and sleep stage included in the model

	Specificity at 90% Sensitivity (%)	Accuracy (%)	AUC	p-value AUC
Baseline model	74.8	81.6	0.914	
Age model	74.3	83.4	0.924	0.36
				0.35
Age + sleep stage model	76.3	82.5	0.925	0.50
Baseline model + LSTM	71.1	83.1	0.914	

Table 4

RESULTS OF TRANSFER LEARNING PER AGE GROUP: TESTED FOR AGE 18–29, 30–39 AND 40–49

Variables	Specificity at 90% Sensitivity (%)	Accuracy (%)	AUC	p-value AUC[a]
Baseline model tested per age group	71.2	79.4	0.894	
- All layers finetuned per age group	63.6	80.4	0.891	0.48
- Last two layers finetuned per age group	66.5	80.4	0.893	0.49
- Last layer finetuned	64.4	81.5	0.890	0.46

[a] p-values are calculated against the baseline model

Table 5

Confusion matrix age model with accuracy in the right bottom corner

		Predicted		
		Normal	**Abnormal**	**Sens/spec**
Actual	Normal	355 (41.7%)	32 (3.8%)	91.7%
	Abnormal	121 (14.2%)	343 (40.3%)	73.9%
	Precision	74.6%	91.5%	82.0%

Table 6

EEG report word analysis of the test set of the age + sleep stage model

		Abnormal EEGs		Normal EEGs	
Category	Words	Ratio of misclassified/ correctly classified	p-value (t-test)	Ratio of misclassified/ correctly classified	p-value (t-test)
Sleep	Sleep, drows*, N1, N2, N2, REM, spindles	1.90	<0.01†	0.96	0.66
Artifacts	Quality, artifact, difficult, nois*, unsatisfactory, myogenic	0.50	0.06	2.80	0.03†
Medication	Medication, medicine, meds, sedat	0.78	0.32	0.22	0.10
Small	Subtle, small, little, slight, minor, modest, limited	1.10	0.78	1.49	0.57
Large	Large, great, clear, apparent, evident, substantial	0.75	0.25	2.14	0.22
Doubt	Probable, maybe, mildly abnormal, possibl*, plausib*	1.93	0.04†	2.35	0.10
Slowing	Theta, delta, slowing	0.71	<0.01†	1.77	<0.01†
Spike	Spike, sharp, wave, discharge	1.18	0.08	1.56	<0.01†
Diffuse	Generalized, diffuse, continuous, frequent, regular	0.71	<0.01†	1.52	<0.01†
Intermittent	Intermittent, rare, infrequent, irregular, occasional, focal	1.15	0.26	1.62	0.24

†significant (<0.05) with a two-tailed T-test testing the misclassified and correctly classified groups. Ratios >1 indicate words associated with misclassification.