Confused or not Confused?

Disentangling Brain Activity from EEG Data Using Bidirectional LSTM Recurrent Neural Networks

Zhaoheng Ni
The Graduate Center, City University of New York
New York, NY 10016, USA
zni@gradcenter.cuny.edu

Ahmet Cem Yuksel
The Graduate Center, City University of New York
New York, NY 10016, USA
ayuksel@gradcenter.cuny.edu

Xiuyan Ni
The Graduate Center, City University of New York
New York, NY 10016, USA
xni2@gradcenter.cuny.edu

Michael I Mandel
Brooklyn College, City University of New York
Brooklyn, NY 11210, USA
mim@mr-pc.org

Lei Xie*
Hunter College, City University of New York
New York, NY 10065, USA
lxie@iscb.org

ABSTRACT

Brain fog, also known as confusion, is one of the main reasons for low performance in the learning process or any kind of daily task that involves and requires thinking. Detecting confusion in a human's mind in real time is a challenging and important task that can be applied to online education, driver fatigue detection and so on. In this paper, we apply Bidirectional LSTM Recurrent Neural Networks to classify students' confusion in watching online course videos from EEG data. The results show that Bidirectional LSTM model achieves the state-of-the-art performance compared with other machine learning approaches, and shows strong robustness as evaluated by cross-validation. We can predict whether or not a student is confused in the accuracy of 73.3%. Furthermore, we find the most important feature to detecting the brain confusion is the gamma 1 wave of EEG signal. Our results suggest that machine learning is a potentially powerful tool to model and understand brain activity.

CCS CONCEPTS

•Computing methodologies → Neural networks; Feature selection; •Applied computing → *Bioinformatics;*

KEYWORDS

Confusion Detection, EEG, LSTM, Machine Learning

ACM Reference format:
Zhaoheng Ni, Ahmet Cem Yuksel, Xiuyan Ni, Michael I Mandel, and Lei Xie. 2017. Confused or not Confused?. In *Proceedings of ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.,* 6 pages.
DOI: http://dx.doi.org/10.1145/3107411.3107513

*The corresponding author

ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.
 ISBN 978-1-4503-4722-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3107411.3107513

1 INTRODUCTION

1.1 Motivation

Brain fog is a constellation of symptoms that include reduced mental acuity and cognition, inability to concentrate and multitask, and loss of short-term and long-term memory. It is well distributed among not only patients with brain diseases but also healthy people [1]. Brain confusion, which is one of the symptoms of brain fog, can reduce people's concentration and cognition. Detecting and preventing brain confusion is very important and has many benefits. When a driver is confused, his/her cognition is reduced. This is very dangerous and can cause serious consequences. Another example is Massive Open Online Course (MOOC), which is an online course aiming at unlimited participation and open access via the web. Although there are several MOOC websites, the format still has shortcoming compared with traditional classes. Valerie et al. [2] showed that the lack of feedback is one of the main problems of student-teacher long distance communication. The students may feel confused about the lecture while the teacher doesn't notice and continues the lecture. If there is a practical approach to detect student's confusion immediately, it will help teachers understand better the students' status and react accordingly.

Electroencephalography (EEG) is an electro-physiological monitoring method to record electrical activity of the brain. In clinical contexts, EEG refers to the recording of the brain's spontaneous electrical activity over a period of time, as recorded from multiple electrodes placed on the scalp. It measures voltage fluctuations resulting from ionic currents within the neurons of the brain. EEG is most often used to diagnose epilepsy, which causes abnormalities in EEG readings. It is also used to diagnose sleep disorders, coma, encephalopathy, and brain death. Our motivation for choosing EEG signals as the data for detecting confusion in people's brains is that EEG signal is continuous and contains some patterns of status transitions. Our hypothesis is that when people are confused, their EEG signal will differ from normal. It is possible to build a model to analyze the continuous data and predict whether the subject is confused or not.

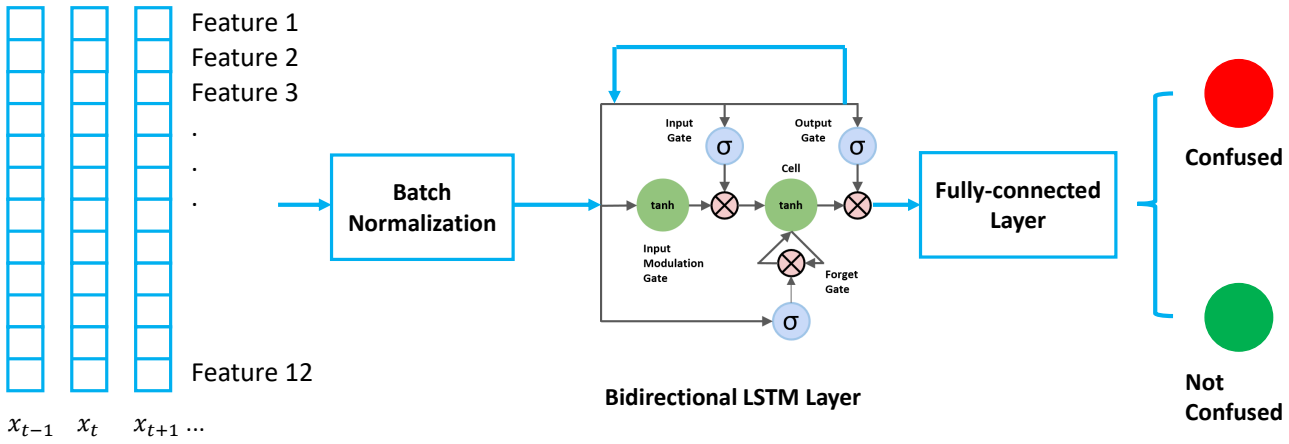

Figure 1: Framework of Bidirectional LSTM model.

1.2 Related Work

There is a general agreement that visual inspection of EEG waveforms patterns can reliably identify driver fatigue or drowsiness. There are many researchers applying machine learning methods to EEG data to accomplish different tasks, such as driver fatigue detection. Yeo et al. [3] used Support Vector Machines (SVMs) to detect the drowsiness of car drivers. Their results showed that extracting features from four EEG frequency bands achieved 99.3% accuracy. Besides drowsiness, Subashi et al. [4] applied SVM classifiers to predict if EEG signals represented epileptic seizures and achieved 100% accuracy. Wang et al. [5] showed the possibility of using EEG data to detect the confusion of students when they watch MOOC videos. They analyzed the EEG data using Gaussian Naive Bayes classifiers. The Gaussian Naive Bayes classifiers achieved a classification accuracy of 57%. The current paper explores methods to improve this confusion classification result on the same dataset.

Recently, deep learning has shown its power on many classification-related tasks compared with traditional machine learning approaches. Boureau et al. [6] proposed a Deep Belief Network (DBN) that can learn a high-level feature based on raw input and can capture higher-order dependencies between observed variables. Hajinoroozi et al. [7] applied DBNs to EEG signals to predict drivers' cognitive states. Classifiers using the features learned by DBNs outperformed those using Principal Component Analysis (PCA) features. Lee et al. [8] introduced convolutional DBNs to learn better feature representations and outperformed machine learning approaches using raw features. Due to the fact that the EEG signal is a time-series, however, detecting events in EEG signals using fixed-length features may be difficult.

Laurent et al. [2] proposed a Hidden Markov Model-based approach for mental state detection in EEG signals. Petrosiana et al. [9] showed that Recurrent Neural Networks can identify early signs of Alzheimer's disease in long-term EEG recordings. Few studies focus on detecting confusion from EEG signals using Deep Neural Networks (DNNs). Since LSTM Recurrent Neural Networks can easily analyze time-series data, the current study applies them to detecting confusion in EEG signal.

We utilize batch normalization, which has been shown to speed training of DNNs. Ioffe et al. [10] proposed a batch normalization layer, which uses mini-batch statistics to standardize features in deep neural networks which can achieve the same accuracy in much less time. Laurent et al. [11] showed that applying batch normalization, to Recurrent Neural Networks leads to a faster convergence of training.

1.3 Problem Statement

Given EEG data from 10 college students, our task is to predict their confusion using machine learning methods. The data is from the "EEG brain wave for confusion" data set, an EEG data from a Kaggle challenge [12]. 10 students were assigned to watch 20 videos, 10 of which were pre-labeled as "easy" and 10 as"difficult". Each video was about 2 minutes long. For "difficult" videos, the two-minute clip was taken abruptly from the middle of a topic to make the videos more confusing.

The students wore a single-channel wireless MindSet EEG device that measured activity over the frontal lobe. The MindSet measures the voltage between an electrode resting on the forehead and two electrodes (one ground and one reference) each in contact with an ear [13]. After each session, the student rated his/her confusion level on a scale of 1-7, where one corresponded to the least confused and seven corresponded to the most confused. These labels were quantized into two classes representing whether the students were confused or not. The two-class label serves as the target of our prediction task.

Since the confusion label is true or false, our problem is a two-class classification problem. In theory, many machine learning approaches can be applied to this task. To take advantage of EEG data's properties, we propose a confusion detection framework using LSTM Recurrent Neural Networks.

Table 1: Features extracted from MindSet. Table from [5]

Features	Description	Sampling rate	Statistic
Attention	Proprietary measure of mental focus	1 Hz	Mean
Meditation	Proprietary measure of calmness	1 Hz	Mean
Raw	Raw EEG signal	512 Hz	Mean
Delta	1-3 Hz of power spectrum	8 Hz	Mean
Theta	4-7 Hz of power spectrum	8 Hz	Mean
Alpha1	Lower 8-11 Hz of power spectrum	8 Hz	Mean
Alpha 2	Higher 8-11 Hz of power spectrum	8 Hz	Mean
Beta1	Lower 12-29 Hz of power spectrum	8 Hz	Mean
Beta 2	Higher 12-29 Hz of power spectrum	8 Hz	Mean
Gamma1	Lower 30-100 Hz of power spectrum	8 Hz	Mean
Gamma2	Higher 30-100 Hz of power spectrum	8 Hz	Mean

2 CONFUSION DETECTION FRAMEWORK

Based on the consideration that these data are continuous in time and the detection sample should not be too long to permit real-time processing, we propose a confusion detection framework as shown in Figure 1. The dataset provides EEG features organized by time. We then use Batch Normalization to normalize the value of each feature to have a mean of 0 and standard deviation of 1. We then train a Bidirectional LSTM model and evaluate its performance using 5-fold cross validation. We also test the contribution of each variable to the model and rank these contributions. The framework is designed as followed:

2.1 Feature Extraction

The EEG data can be downloaded from the Kaggle website [12], which provides open source data for various challenges. In the EEG dataset, 10 college students are asked to wear a wireless single-channel MindSet EEG device [13] that measured activity over the frontal lobe and to watch 10 2-minute long videos. The MindSet extracted the features that are shown in Table 1. "Attention" measures the mental focus of the student, and "Meditation" measures calmness. "Raw" is the average of the original EEG signals. The following features are values in different frequency regions of the power spectrum. The sampling frequency for the features extracted from the MindSet is 2 Hz. For each sample point, there are 14 features extracted from EEG signals, shown in Table 2.

"Subject id" ranges from 0 to 9, representing the subject of each recording, "video id" is the same for videos. We don't use them as features in our model. In the feature representations, we also have power spectrum for specific frequencies, which are all continuous data.

For each subject watching a video, features are extracted at a sampling frequency of 2 Hz. The final features are truncated to around one-minute long. So there are around 120 sample features for each data point. For models the only accept fixed-length features, we took the minimum number of samples (112 samples at last) and concatenated all the time steps to create a single feature vector. For the LSTM, the length of the time-series data is 112, each with a 12-dimension feature. In the end there are 100 data points, each with 112×12 features in total.

Table 2: Features with index.

F1	Attention
F2	Meditation
F3	Raw
F4	Delta
F5	Theta
F6	Alpha1
F7	Alpha2
F8	Beta1
F9	Beta2
F10	Gamma1
F11	Gamma2
F12	Predefined Label
F13	Subject id
F14	Video id

2.2 RNN-LSTM and Bidirectional LSTM

In RNNs, back propagation flows through many layers, passing through many stages of multiplication. During the training process, error messages flowing backward in time tend to either blow up, which may lead to oscillating weights, or vanish in which case they become too small to provide a learning signal.

Exploding gradients can be mitigated via truncation or squashing. Vanishing gradients are more difficult to fix. The Long Short-Term Memory RNN (LSTM) addresses this problem by introducing memory units to RNNs. The memory units help preserve the error signal so that it is large enough to be back propagated through time and layers, thereby opening a channel that links remote causes and effects. The architecture of the LSTM is shown in Figure 2.

Given an input sequence $X = (x_1, x_2, x_3, \ldots, x_T)$, the hidden state in time t in RNN is defined as followed:

$$h_t = \Phi(W_h h_{t-1} + W_x x_t + b), \quad (1)$$

where $W_h \in \mathbb{R}^{d_h \times d_h}$, $W_x \in \mathbb{R}^{d_x \times d_x}$, $b \in \mathbb{R}^{d_h}$.

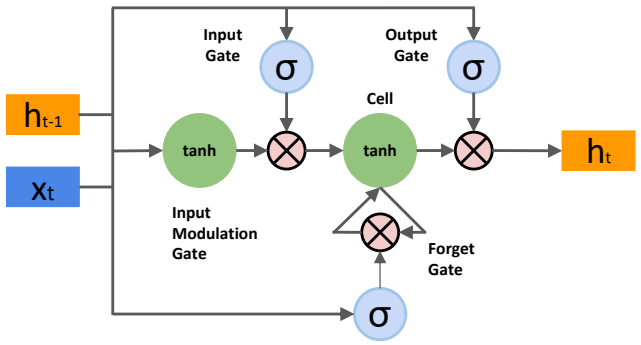

Figure 2: Architecture inside LSTM cell.

The architecture of the LSTM is defined as followed:

$$\begin{pmatrix} \tilde{f}_t \\ \tilde{i}_t \\ \tilde{o}_t \\ \tilde{g}_t \end{pmatrix} = W_h h_{t-1} + W_x x_t + b, \quad (2)$$

$$c_t = \sigma(\tilde{f}_t) \odot c_{t-1} + \sigma(\tilde{i}_t) \odot tanh(\tilde{g}_t), \quad (3)$$

$$h_t = \sigma(\tilde{o}_t) \odot tanh(c_t), \quad (4)$$

where $W_h \in \mathbb{R}^{d_h \times 4d_h}$, $W_x \in \mathbb{R}^{d_x \times 4d_x}$, σ is the logistic function, and the $\odot$ operation is the Hadamard product.

LSTM RNNs can learn long-term temporal dynamics that traditional RNNs cannot. Using its memory cells, it learns to forget previous memories and considering the current input, determines how much of the memory to be transferred to the next hidden state. In our case, since EEG features arrive over time, the LSTM can incorporate context information across time to improve performance.

While LSTMs predict the current output based on previous inputs, bidirectional LSTMs predict the current output based on both the past and the future. The basic idea is to predict the future based on the past and predict the past based on the future, then take the average of these two outputs as the finial output. In this way both future and past context information can be utilized to improve performance.

In our framework, we set the Bidirectional LSTM layer to have 50 neural units. The activation function in the LSTM layer is *tanh*. After the LSTM layer, the hidden states are fed into a fully connected layer with a sigmoid activation function, which produces output value between 0 and 1.

2.3 Batch Normalization

Training Deep Neural Networks is complicated by the fact that the distribution of the inputs to each layer changes during training, as the parameters of the previous layers change [14]. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously difficult to train models with saturating nonlinearities.

Recently, Ioffe et al. [10] proposed a Batch Normalization method which can be built as a sub-architecture into a model. Batch Normalization allows us to use much higher learning rates and be less careful about initialization. They showed that adding Batch Normalization to a state-of-the-art image classification model yields a substantial speedup in training. By further increasing the learning rates, removing Dropout, and applying other modifications afforded by Batch Normalization, their model matches the previous state of the art in only a small fraction of the number of training steps and then beats the state of the art in single-network image classification. Batch Normalization is defined as:

$$BN(x, \gamma, \beta) = \beta + \gamma \frac{x - \hat{\mathbb{E}}(x)}{\sqrt{\hat{Var}(x) + \epsilon}} \quad (5)$$

where x is the vector that needs to be normalized. $\hat{\mathbb{E}}(x)$ and $\hat{Var}(x)$ are the expectation and variance of the current mini-batch of x, respectively, ϵ is a constant added to the mini-batch variance for numerical stability, and β is a parameter to shift the normalized value.

In our framework, we normalize the training data in a feature-wise fashion (i.e., each feature dimension is normalized to have a mean of 0 and standard deviation of 1 across each batch of samples). The batch size is set to 20, which corresponds to our test data size. After Batch normalization, we put the normalized features into our Bidirectional LSTM model.

3 EXPERIMENTS

3.1 Baseline Models

To evaluate our framework's performance, we designed several baseline machine learning approaches, listed in Table 3.

Table 3: Baseline classification methods for classifying confusion.

Baseline Classification Methods
SVM (linear kernel)
SVM (rbf kernel)
SVM (sigmoid kernel)
K-Nearest Neighbors
Convolutional Neural Network
Deep Belief Network
RNN-LSTM

We apply three SVM classifiers using different kernel functions. We use grid search to tune the parameters C ranging in (1, 10, 100, 1000) and γ ranging in (10^{-3}, 10^{4}) for each kernel, respectively. We also apply a K Nearest Neighbor classifier as another baseline method. We use different K parameter values ranging from 2 to 5 and choose the highest accuracy as the final result. To compare the results of different neural networks, we use Convolutional Neural Network, Deep Belief Network, and a single-layer LSTM Recurrent Neural Network (RNN-LSTM) to classify the EEG data.

3.2 Variable Selection

To test which feature of the EEG dataset contributes the most to our model, we propose a variable selection method to find the most important feature in our Bidirectional LSTM model. Not putting all the features into the model, instead we leave one single feature out. Then we run the experiments with the remaining features. After we get the average accuracy across cross validation folds, we rank the accuracy from lowest to highest, providing the feature importance from highest to lowest.

4 RESULTS

4.1 Cross-Validation Results

To evaluate the models, we perform 5-fold cross validation. The result is shown in Table 4.

Table 4: Average accuracy for 5-fold cross validation.

Classification methods	Accuracy(%)
SVM (linear kernel)	67.2
SVM (rbf kernel)	51.3
SVM (sigmoid kernel)	51.0
K-Nearest Neighbors	51.9
Convolutional Neural Network	64.0
Deep Belief Network	52.7
RNN-LSTM	69.0
Bidirectional LSTM	73.3

The results show that the Bidirectional LSTM achieves the best performance compared to the other methods. Due to the nature of the dataset, the Deep Belief Network cannot approximate the data distribution with only 100 data points. Neither can the Convolutional Neural Network or K-Nearest Neighbors classifier. The SVM with linear kernel performs similarly to the RNN-LSTM, and outperforms the SVMs with more complex kernels. This result implies that the feature space is almost linearly separable. Classifying the data in this linear space is better than other spaces, which may cause overfitting.

Then we evaluated the robustness of the RNN-LSTM model and the Bidirectional LSTM model by analyzing the accuracy of each iteration of cross validation. The results are shown in Figures 3 and 4

The accuracy across the five folds of the RNN-LSTM model varies from 60 percent to 85 percent, while that of the Bidirectional LSTM model varies from 71 percent to 74 percent. These results show that the Bidirectional LSTM model not only outperforms all the other methods, but also is consistent.

4.2 Variable Selection

To find the feature that makes the most contribution to the model, we ran 12 experiments, each leaving out one feature. The accuracy of each of these classifiers is shown in Figure 5. From the rank among accuracies, we can see that losing feature 10, which is the gamma-1 feature, decreases the accuracy the most. The beta-2 and attention features also lead to large decreases in performance.

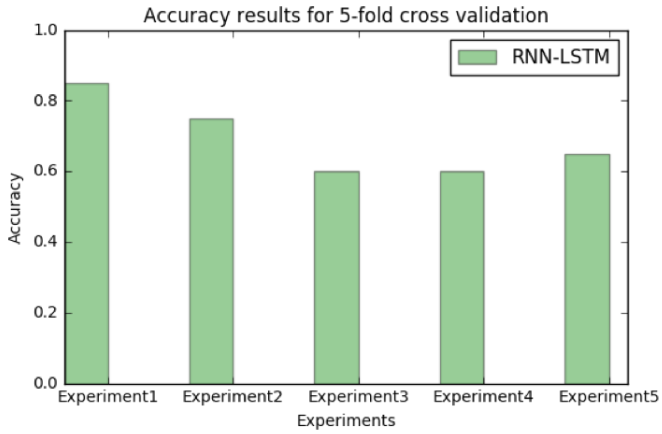

Figure 3: Accuracy variation of RNN-LSTM model.

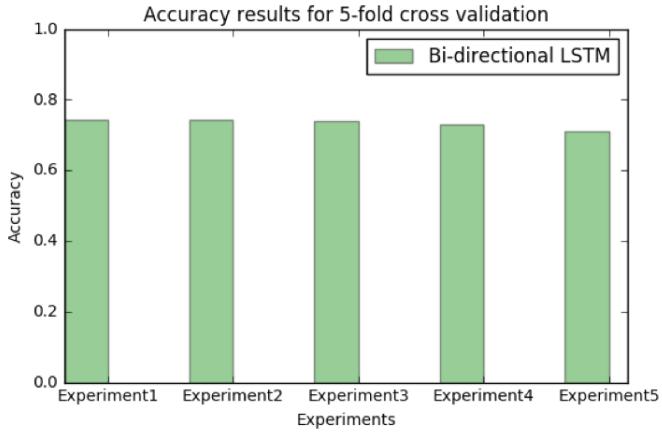

Figure 4: Accuracy variation of Bidirectional LSTM model.

Based on the rank of the features, we can select the most important features for the confusion detection system. In this way we can still maintain high accuracy and also make it possible for the system to detect confusion in real time.

We also found that the pre-defined video label contributed the least to our Bidirectional LSTM model's accuracy. The subjective judgment of video difficulty of the experiment designer is different from that of the students, a very interesting point. This can help teachers identify topics that students don't understand, while the teachers may think the class is easy for students.

For this specific dataset, it is hard to build a model from so few examples. Hence for deep neural networks such as the DBN and CNN it is hard to tune the parameters perfectly and easy to overfit. However, for the LSTM, though different time steps (i.e. feature for each 0.5 second) input to the LSTM share the same weights in the neural network, the forget gate can learn how to make use of previous hidden states. Bidirectional LSTMs make use of sequential information in both directions and learn a better representation. Adding context information helps us build a more robust and accurate model.

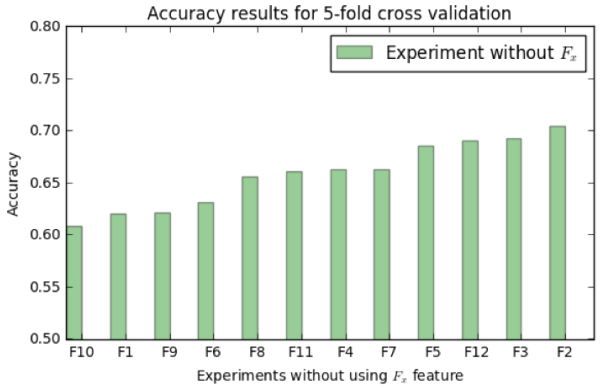

Figure 5: Accuracy without specific feature from 12 features. Ranked from lowest accuracy to highest.

5 CONCLUSIONS

We have proposed a Bidirectional LSTM Recurrent Neural Network framework to detect students' confusion when watching online course videos. The accuracy achieved by our model is higher than other machine learning approaches including a single-layer RNN-LSTM model and achieves the state-of-the-art result. The architecture of the Bidirectional LSTM model takes advantage of time-series features and helps improve performance. By analyzing the contribution of each feature to the model, we find the "gamma-1" and "attention" features are the most important in this task. We plan to validate our model on a larger EEG dataset. We also plan to apply our Bidirectional LSTM model on other EEG-related tasks, such as driver drowsiness detection.

ACKNOWLEDGMENTS

We appreciate reviewers constructive commends. This research was supported under National Institute of Health Grant R01LM011986.

REFERENCES

[1] Theoharis Constantin Theoharides, Julia M Stewart, and Erifili Hatziagelaki. Brain "fog," inflammation and obesity: key aspects of neuropsychiatric disorders improved by luteolin. *Frontiers in neuroscience*, 9:225, 2015.
[2] Laurent Vézard, Pierrick Legrand, Marie Chavent, Frédérique Faïta-Aïnseba, and Leonardo Trujillo. EEG classification for the detection of mental states. *Applied Soft Computing*, 32:113–131, 2015.
[3] Mervyn VM Yeo, Xiaoping Li, Kaiquan Shen, and Einar PV Wilder-Smith. Can SVM be used for automatic EEG detection of drowsiness during car driving? *Safety Science*, 47(1):115–124, 2009.
[4] Abdulhamit Subasi and M Ismail Gursoy. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 37(12):8659–8666, 2010.
[5] Haohan Wang, Yiwei Li, Xiaobo Hu, Yucong Yang, Zhu Meng, and Kai-min Chang. Using EEG to improve massive open online courses feedback interaction. In *AIED Workshops*, 2013.
[6] Y-lan Boureau, Yann L Cun, et al. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2008.
[7] Mehdi Hajinoroozi, Tzyy-Ping Jung, Chin-Teng Lin, and Yufei Huang. Feature extraction with deep belief networks for driver's cognitive states prediction from EEG data. In *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*, pages 812–815. IEEE, 2015.
[8] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.
[9] AA Petrosian, DV Prokhorov, W Lajara-Nanson, and RB Schiffer. Recurrent neural network-based approach for early recognition of alzheimer's disease in EEG. *Clinical Neurophysiology*, 112(8):1378–1387, 2001.
[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
[11] Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. Recurrent batch normalization. *arXiv preprint arXiv:1603.09025*, 2016.
[12] Haohan Wang. EEG brain wave for confusion, 2016. Online: https://www.kaggle.com/wanghaohan/eeg-brain-wave-for-confusion.
[13] NeuroSky. NeuroSky's eSense meters and detection of mental state. 2009.
[14] César Laurent, Gabriel Pereyra, Philémon Brakel, Ying Zhang, and Yoshua Bengio. Batch normalized recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2657–2661. IEEE, 2016.
[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.