2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Importance of the Features of Event-Related Potentials Used for a Machine Learning-Based Model Applied to Single-Trial Data during Oddball Task

Naohito Yoshioka, Nobuyuki Araki, and Mieko Ohsuga

Abstract— **In this study, a method for assessing the human state and brain-machine interface (BMI) has been developed using event-related potentials (ERPs). Most of these algorithms are classified based on the ERP characteristics. To observe the characteristics of ERPs, an averaging method using electroencephalography (EEG) signals cut out by time-locking to the event for each condition is required. To date, several classification methods using only single-trial EEG signals have been studied. In some cases, the machine learning models were used for the classifications; however, the relationship between the constructed model and the characteristics of ERPs remains unclear. In this study, the LightGBM model was constructed for each individual to classify a single-trial waveform and visualize the relationship between these features and the characteristics of ERPs. The features used in the model were the average values and standard deviation of the EEG amplitude with a time width of 10 ms. The best area under the curve (AUC) score was 0.92, but, in some cases, the AUC scores were low. Large individual differences in AUC scores were observed. In each case, on checking the importance of the features, high importance was shown at the 10-ms time width section, where a large difference was observed in ERP waveforms between the target and the non-target. Since the model constructed in this study was found to reflect the characteristics of ERP, as the next step, we would like to try to improve the discrimination performance by using stimuli that the participants can concentrate on with interest.**

I. Introduction

Event-related potentials (ERPs) are used to assess the human state [1]-[3] and build brain-machine interfaces (BMI) [4][5]. ERPs are small components that appear to be accompanied by a particular event and are usually difficult to observe without electroencephalography (EEG) processing owing to the basic rhythm of EEGs and external noise. Cutting out EEG signals triggered by events and averaging for the same recording conditions cancels EEG signals unrelated to the event and noise. However, observing ERPs with a few addition-averaging (ideally a single trial) is desirable because the averaging process reduces the efficiency of human state estimation and BMI. The oddball task is a relatively simple technique used in ERP experiments that randomly provides target and non-target stimuli to the experimental participants to discriminate the target, and a large ERP component approximately 300 ms can be obtained when the frequency of the target is low. EEG signals are obtained during many trials for target and non-target stimuli, and the differences in the characteristic components of ERPs between target and non-target trials were investigated. The method for classifying single-trial EEG signals into targets or non-targets has been studied. Typical examples are prediction models using machine learning, such as support vector machines (SVMs) [6], or deep learning, such as convolutional neural networks (CNNs) [7]. In both cases, the in-model structure of the features used for model construction cannot be determined; therefore, obtaining a hint to improve the models is difficult.

The objective of this study is to construct a prediction model based on ensemble learning and to clarify the model structure of features in the constructed model.

II. Experiment

The experiment was executed with the permission of the president of Osaka Institute of Technology in accordance with the report of the Life Science Ethics Committee of Osaka Institute of Technology (No. 2018-12-2). The participants were healthy adults who provided written informed consent. Ten males and ten females aged between 24 and 61 years (average age: 41.3 years) participated.

A. Experimental method

The oddball task using visual stimuli was presented in a soundproof room. The participants were seated in front of the display. The height of the participant's eye point and the center of the screen were matched, and the distance between the eye point and display was maintained at 70 cm. A box installed with two buttons was prepared at the participant's hand to respond to the discrimination results. EEG and electrooculography (EOG) signals were measured using a general-purpose biological amplifier (PolymateV, Miyuki Giken Co., Ltd.) during the experiment. The time constant was 3 s, and the sampling rate was 1000 Hz. EEGs were recorded using active electrodes (AP-C151(A)-015, Miyuki Giken Co., Ltd.) placed at 11 locations (F3, Fz, F4, C3, Cz, C4, T5, P3, Pz, P4, T6) based on the International 10-20 system. EOGs were recorded using small bioelectrodes (NT-211U, Nihon Kohden Co., Ltd.) attached to the outsides of both eyes, as well as above and below the dominant eye. The reference electrodes were placed on both earlobes, and the ground was attached to the forehead. To obtain data to investigate the inclusion of EOGs in EEGs, an eye-movement trial was executed before the oddball task while the participants were required to blink and to move their eyes left and right, as well as up and down.

Naohito Yoshioka is with the Graduate School of Robotics and Design, Osaka Institute of Technology, Osaka 5308568 Japan, and Yanmar Holdings Co., Ltd., Shiga 5218511 Japan (e-mail: d1d19r01@st.oit.ac.jp, naohito_yoshioka@yanmar.com).

Nobuyuki Araki is with Yanmar Holdings Co., Ltd., Shiga 5218511 Japan (e-mail: nobuyuki_araki@yanmar.com).

Mieko Ohsuga is with faculty of Robotics and Design, Osaka Institute of Technology, Osaka 5308568 Japan (e-mail: mieko.ohsuga@oit.ac.jp).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

B. *Experimental task*

Two target stimuli and seven non-target stimuli, shown in Fig. 1, were used in the oddball task. A total of 162 stimuli were presented 18 times for each stimulus. The presentation program comprised 18 blocks, in which 9 types of stimuli were randomly arranged in one block. The stimuli for 0.5 s and the gaze point (+ mark) for 0.7 s were presented repeatedly.

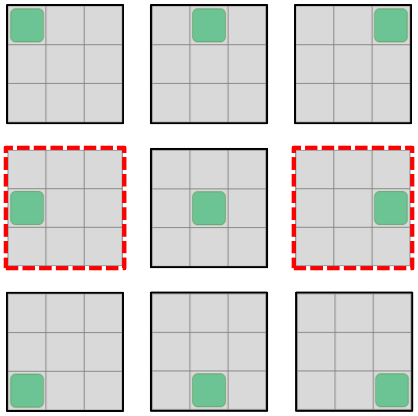


Figure 1. Visual stimuli. The target stimuli are designated by the red dashed lines.

C. *Preprocessing method for EEGs*

The EEGs were analyzed using EEGLAB [8][9]. They were processed using a notch filter (60 Hz) and a band-pass filter (1–30 Hz). The artifacts of the EOG components were removed using the data recorded before the experiment using the independent component analysis algorithm. The onset of the stimulus was set as zero on the time axis, and EEG signals from −200 ms to 600 ms were removed. The baseline of the EEG signals was corrected using the averaged amplitude from −200 ms to 0 ms. The EEG signals at four locations (C3, Cz, C4, and Pz) were averaged because the characteristic peaks of ERP in the oddball task were observed clearly around these four locations in our previous research [1].

III. Model Construction

The binary classifications for EEGs obtained in the trial of target stimuli or non-target stimuli were performed using LightGBM models.

A. *LightGBM model*

The LightGBM model is a type of ensemble learning method [10]. It is a gradient-boosting method with a decision tree. One of the advantages of LightGBM is faster training speed [11] because an efficient learning process is adopted in the model [12]. In addition, it can visualize the feature importance. With these features, checking whether the designed features were used, as expected in the model, is possible.

B. *Dataset*

Preprocessed EEG signals were split into the first nine blocks and the latter nine blocks. The first half was used as the training data for constructing the LightGBM models, and the latter half was used as test data for the binary classification of the single-trial EEG signals. Five models using EEG data averaged 1–5 times were constructed using the dataset shown in Table I for each participant. After averaging the signals for all combinations in each class, the non-target dataset for training was randomly selected and reduced in number to match the same size as the target dataset. As the single-trial datasets were small, the class-weighted method was adopted, which performed better than the undersampling method.

TABLE I. The Dataset for Training

Number of addition-averaging	**Number of EEG signals in the training dataset**	
	Target	***Non-target***
Single trial (St)	18	63
2	153 ($_{18}C_2$)	1953* ($_{63}C_2$)
3	816 ($_{18}C_3$)	39711* ($_{63}C_3$)
4	3060 ($_{18}C_4$)	595665* ($_{63}C_4$)
5	8568 ($_{18}C_5$)	7028847* ($_{63}C_5$)

* Undersampled before parameter tuning and training.

C. *Feature design*

After preprocessing and averaging, features were designed using the EEG signals from 0 ms to 600 ms. ERPs during oddball have characteristic peaks. To capture the features of the peaks, we focused on the mean amplitude and the standard deviation for sections divided by a time width of 10 ms. The mean amplitude indicates peak height, and the standard deviation captures the steepness of change before and after the peak. The 120 sections, including 10 sampling points, were obtained for a 600-ms period; that is, 120 feature sets were used for the model (Fig. 2).

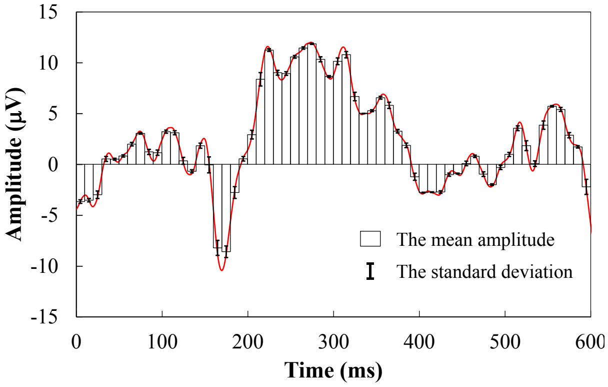


Figure 2. Mean amplitudes and the standard deviations were obtained for every 10-ms time width.

D. *Parameter tuning*

Non-target datasets of the training datasets were randomly undersampled. Without the case of a single trial, undersampling was repeated 10 times, and 10 datasets were obtained for each participant and the number of addition-averaging. All undersampled datasets were split into training and validation subsets for each class. The ratio of the number of training sub-data to verification sub-data was 2:1. The hyperparameter of each dataset was tuned using the combination method optuna module [13]-[15] and stratified k-fold cross-validation imported from the scikit-learn module [16]. The combinations of the undersampled dataset and the hyperparameters that recorded the best averaged AUC score for the validation subset were adopted for the model construction for each participant and the number of addition-averaging.

IV. Classification Result

Using the constructed model, the single-trial EEG signals in the latter nine blocks were classified as target or non-target. Table II shows the classification results for all cases. The best score was 0.92. Some can be classified with very high AUC scores, while others cannot be classified well. Figure 3 shows the average and standard deviation of the AUC scores of all participants for each number of addition averages. The effect of the number of additive averages was significant [F (4, 19) = 9.37, p < 0.001]. The result of a Tukey–Kramer test demonstrated that the AUC score of a single trial was lower than that of the other models. Overall, AUC scores were not good in the cases of single-trial models, and they were improved in two-times averaging models. Even if the number of addition-averaging increases further, the performance does not improve.

V. Consideration

In some participants, the two-times averaging models had high AUC scores. To clarify this reason, the feature importance of LightGBM was compared to the ERP waveform because features were designed based on the characteristic peaks of the ERP waveform.

For each model, the feature importance was divided by the maximum value of the 120 features. For each participant, ERP waves were observed by averaging the EEG signals for 18 target cases and 63 non-target cases. Figure 4 shows contour maps of the feature importance and ERP waveforms of the 2nd and 20th participants. In both cases, as the number of addition-averaging increases, the model focuses more strongly on the time width, where there is a clear difference between the two classes. Many addition-averaging models could not evaluate noise included in single-trial EEG signals and did not obtain a good AUC score. On the contrary, the two-times addition-averaging models used almost all features evenly, so the AUC score improved. However, the AUC score of many addition-averaging models did not drop significantly due to excessive focus on a specific time width, regardless of whether the score was good or bad for each participant. It may be better to consider that the noise-filled single-trial EEG signals were similar between the two classes. The grand-averaged waveform is an ideal single-trial ERP waveform. For example, if pulse noise is mixed in the non-target waveform at approximately 300 ms, as shown in Fig. 4(b), for any model to classify between the two classes of waveforms is impossible. To apply to BMI, considering a method for inducing the difference between the two classes rather than improving the model may be a better option. The images shown in Fig.1 are very similar. Azizian reported that the difference in ERP waveforms was reduced when the physical characteristics of the target and non-target images were similar [17]. Therefore, it is expected that if images with different spatial characteristics are used, classifying them as No.02, which demonstrated the best performance, is possible. In addition, the magnitude of the difference in the ERP amplitude in the target and non-target regions is affected by concentration during the task. It is also necessary to maintain concentration during a task, such as using stimuli wherein participants are interested.

TABLE II. The Summary Sheet of AUC Scores

Participants	**AUC scores for number of addition-averaging**				
	St	***2***	***3***	***4***	***5***
No. 01	0.72	0.76	0.70	0.73	0.71
No. 02	0.83	0.92	0.90	0.88	0.87
No. 03	0.68	0.76	0.71	0.61	0.65
No. 04	0.66	0.65	0.61	0.59	0.59
No. 05	0.77	0.86	0.86	0.87	0.86
No. 06	0.56	0.72	0.68	0.69	0.66
No. 07	0.50	0.52	0.45	0.45	0.47
No. 08	0.48	0.65	0.62	0.67	0.67
No. 09	0.54	0.60	0.62	0.61	0.64
No. 10	0.56	0.65	0.62	0.64	0.66
No. 11	0.54	0.70	0.72	0.77	0.73
No. 12	0.52	0.63	0.57	0.58	0.56
No. 13	0.46	0.55	0.58	0.60	0.61
No. 14	0.50	0.45	0.46	0.49	0.47
No. 15	0.39	0.56	0.48	0.43	0.44
No. 16	0.73	0.76	0.75	0.76	0.76
No. 17	0.59	0.55	0.57	0.51	0.51
No. 18	0.72	0.77	0.73	0.72	0.70
No. 19	0.60	0.69	0.70	0.70	0.70
No. 20	0.43	0.53	0.46	0.44	0.44

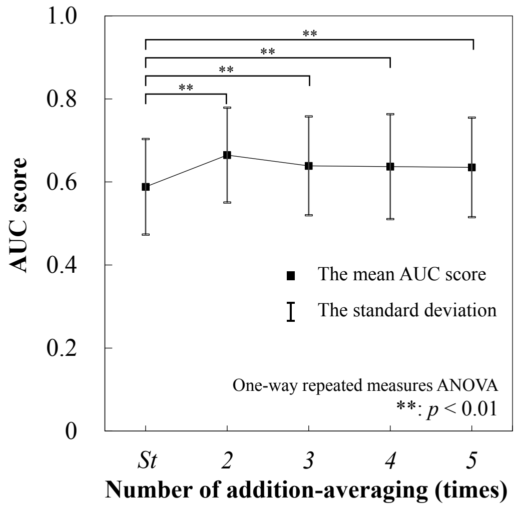
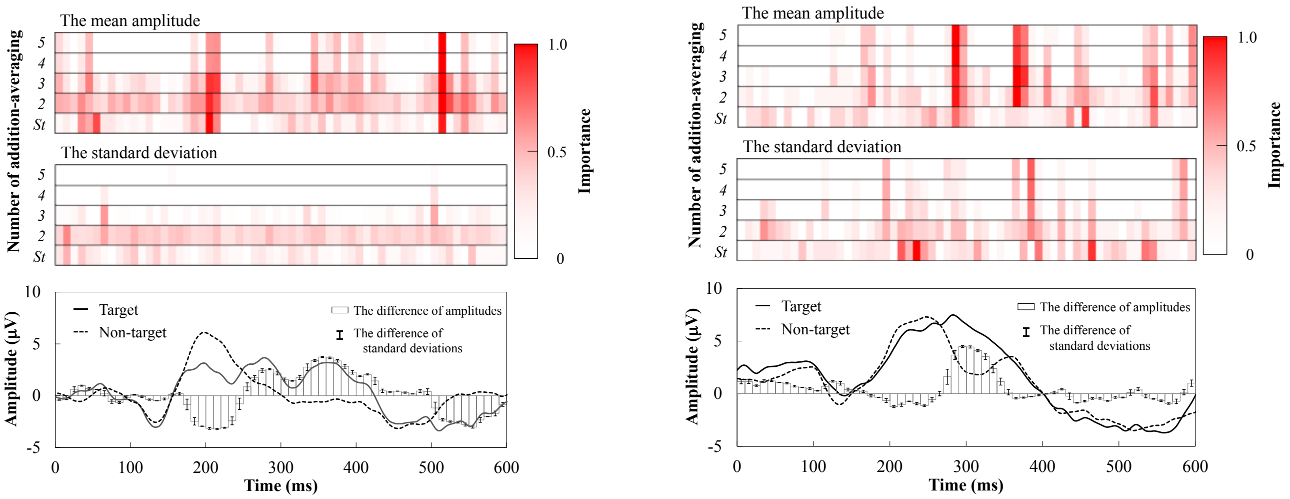


Figure 3. Average and standard deviations of the AUC score



(a) Participants: No.02

(b) Participants: No.20

Figure 4. Feature importance and averaged-ERP waveform. The bar graphs and error bars in the background of the waveforms are the differences of amplitude and standard deviation between the target and the non-target for each time width.

VI. Conclusion

Based on the classification of the single-trial EEG signals using LightGBM models, the conclusions are as follows:

- The structure of features inside model, which could not be visualized by the SVM and CNN models, could be visualized by the LightGBM models. They expressed the characteristic features of ERP waveforms.
- The best AUC score for classification was 0.92, which was very high. The best averaged AUC score for all participants was two-times addition-averaging models. They emphasized some features and used almost all other features evenly in the cases of both participants with good and bad AUC scores. This was effective for the classification of noise-filled single-trial EEG signals.
- In addition to improving the models, a study of the oddball task using clearly different stimuli should be conducted involving stimuli in which participants are interested.

One of the limitations of this study is that we focused only on the LightGBM model in order to quickly reveal the relationship between the feature importance and ERP features. In the future, other algorithms (SVM, CNN, etc.) should be tried and compared in order to build models with better AUC scores.

References

[1] N. Yoshioka, N. Araki, and M. Ohsuga, “Evaluation method for the gap between user-expected HMI layout and actual layout,” *Trans. Soc. Automot. Eng. Japan*, vol. 50, no. 6, Nov. 2019, pp. 1659–1664 (in Japanese).

[2] W. Fruehwirta, G. Dorffner, S. Roberts, M. Gerstgrasser, D. Grossegger, R. Schmidt, P. Bianco, G. Ransmayr, H. Garn, M. Waser, and T. Benke, “Associations of event-related brain potentials and alzheimer’s disease severity: a longitudinal study,” *Neuropsychopharmacol. Biol. Psych.*, vol. 92, 2019, pp. 31–38.

[3] N. Yoshioka, T. Kimura, Y. Shu, T. Okamatsu, N. Araki, and M. Ohsuga, “Evaluation of the tiller switch layout of a tractor using eye-fixation related potentials,” *42nd Annual Int. Conf. IEEE Eng. Med. Biol. Soc.*, Montreal, QC, Canada, 2020.

[4] R. Hasegawa, Y. Nakamura, Y. Hasegawa, and H. Sawahata, “Neural prediction of the target "to BUY" or "NOT to BUY" by the ERP- based Cognitive BMI,” *5th Int. Symp. Affective Sci. Eng.*, Tokyo, Japan, 2019.

[5] R. Chavarriaga and J. Millán, “Learning from EEG error-related potentials in noninvasive brain-computer interfaces,” *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 18, no. 4, 2010, pp. 381–388.

[6] A, H. Neuhaus, F, C. Popescu, J. Rentzsch, and J. Gallinat, “Critical evaluation of auditory event-related potential deficits in schizophrenia: evidence from large-scale single-subject pattern classification,” *Schizophrenia Bulletin*, vol. 40, no. 5, 2014, pp. 1062–1071.

[7] I. Sturm, S. Bach, W. Samek, and K, R. Müller, “Interpretable deep neural networks for single-trial EEG classification,” *J. Neurosci. Methods*, vol. 274, 2016, pp. 141–145.

[8] A. Delorme, “EEGLAB: An open-source toolbox for analysis of single trial EEG dynamics including independent component analysis,” *J. Neurosci. Methods*, vol. 134, 2004, pp. 9–21.

[9] Swartz Center for Computational Neuroscience, EEGLAB, https://sccn.ucsd.edu/eeglab/, (accessed: Aug 03, 2018)

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” *Proc. 31st Conf. Neural Informat. Proc. Systems*, pp. 3148–3156, Long Beach, CA, USA, 2017.

[11] Microsoft Corporation, Lessons Learned From Benchmarking Fast Machine Learning Algorithms, https://docs.microsoft.com/ja-jp/archive/blogs/machinelearning/lessons-learned-benchmarking-fast-machine-learning-algorithms, (accessed: Oct 25, 2020)

[12] H. Shi, “Best-first decision tree learning,” Master Thesis, University of Waikato, 2007.

[13] A hyperparameter optimization framework, https://optuna.readthedocs.io/en/stable/index.html, (accessed: Dec 08, 2020)

[14] J. Bergstra, R. Bardenet, Y. Bengio, and B. K´egl, “Algorithms for hyper-parameter optimization,” 2011.

[15] J. Bergstra, D. Yamins, and D. D. Cox, “Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures,” *Proc. 30th Int. Conf. Machine Learning*, Atlanta, Georgia, USA, 2013.

[16] scikit-learn Machine Learning in Python, https://scikit-learn.org/stable/, (accessed: Dec 08, 2020)

[17] A. Azizian, A. L. Freitas, T. D. Watoson, and N. K. Squires, “Electrophysiological correlates of categorization P300 amplitude as index of target similarity,” *Biol. Psychol.*, vol. 71, 2006, pp. 278–288.