ICIC Express Letters
Part B: Applications
Volume **12**, Number **11**, November **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 1003–1009

DETERMINING THE BEST FEATURE FOR IDENTIFYING THE IMAGINED WORD BASED ON EEG SIGNAL USING FEATURE IMPORTANCE SCORE METHOD

Efy Yosrita[1,*], Yaya Heryadi[1], Lili Ayu Wulandhari[2] and Widodo Budiharto[2]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[2]Department of Computer Science, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
*Corresponding author: efy.yosrita@binus.ac.id
{ yayaheryadi; wbudiharto }@binus.edu; lwulandhari@binus.ac.id

Received February 2021; accepted May 2021

Abstract. *The aim of this study is to select the best features of EEG signal, by investigating the AdaBoost feature importance score measure as a means to find a ranking of important features which can improve the classifier performance for recognizing the imagined speech of 8 Indonesian words, i.e., makan (eat), minum (drink), lapar (hungry), haus (thirsty), senang (happy), sedih (sad), sakit (sick) and toilet (toilet). The EEG signal was recorded from 11 healthy students, 7 men and 4 women, using Emotiv epoch and Emotiv Pro. Feature importance score was applied to AdaBoost model. Our research showed that the top ten features based on feature importance score ranking of AdaBoost model were T7_GAMMA, T7_THETA, P7_HIGH BETA, P8_GAMMA, P8_HIGH BETA, F3_GAMMA, F3_HIGH BETA, T7_HIGH BETA, P7_GAMMA and FC5_THETA, with the resulting accuracy 75%, precision 80% and sensitivity or recall 84%.*

Keywords: Feature importance score, AdaBoost, Confusion matrix, EEG, Feature selection

1. **Introduction.** Imagined speech is one of the research fields of speech recognition based on EEG signals [1-3]. Imagined speech refers to the activity of imagining a word without sound production or moving the muscles around the lips [4]. The imagined speech research is divided into 3, i.e., vowel imagination [5-7], syllable imagination [8,9] and word imagination [10-12]. Some of the stages commonly carried out in imagined speech research include acquisition, preprocessing, feature extraction, feature selection and classification [13-17]. Feature selection is one step in developing a predictive model [18] in which the purpose is to define the best feature for the model by reducing the number of input variables [19] and still big challenge for a successful signal classification [20]. Feature selection and feature importance method are used to see a new point of view of the data to be explored with algorithm modelling [21]. Feature reduction is an important issue [22] and one of processes in machine learning that can reduce the complexity of space [23] with retaining the variable of information [24], therefore making the model easy to interpret [25] and to improve the efficacy of the classifier [26-29]. Many feature importance scores are specific for a type of data [30]. The aim of this study is to select the best features of EEG signal by investigation of the AdaBoost feature importance score measure as a means to find a ranking of important features for recognizing the imagined word of 8 Indonesian words, i.e., makan (eat), minum (drink), lapar (hungry), haus (thirsty), senang (happy), sedih (sad), sakit (sick) and toilet (toilet). The paper is organized as follows: Section 2

DOI: 10.24507/icicelb.12.11.1003

describes previous research on feature selection method, Section 3 explains about method that is used in this study, Section 4 presents experimental results, and Section 5 concludes the paper.

2. Previous Research. Several studies on feature selection have been carried out, including the research conducted by Ma et al. regarding the use of the hybrid filter-wrapper technique for the feature selection approach [31] and a further research on the comparison of RFE-RF, RFE-SVM and Bayesian Model Averaging (BMA) to select the best predictor by Rumao [32]. In the same year, Yang et al. created a high-dimensional EEG feature formation by extracting several features [33]. Rahman et al. used Rényi min-entropy to perform feature selection [34]. Other studies examined the utilization of PCA to reduce signal dimensions and select the best features by Tiwari and Chaturvedi [35] and the use of the Shapley value method for feature impact analysis [36]. The next research is a survey conducted by Baig et al. regarding filtering techniques for channel selection in EEG motor images [37]. Research on feature importance has been carried out in recent years, e.g., feature selection based on feature importance by Ellies-Oury et al. [38], measurement of variable or feature importance based on the ExtraTree model by Hallett et al. [39] and analysis of some feature importance methods by Wei et al. [40]. Other researchers analyzed variable/feature importance in imbalanced data [41] and proposed estimation method for efficiency of developing machine learning models based on nonparametric variable importance and utilization of clustering use binary decision trees (CUBT) to define feature importance [42].

3. Research Method. The process of EEG signal's feature selection based on feature importance score is shown in Figure 1.

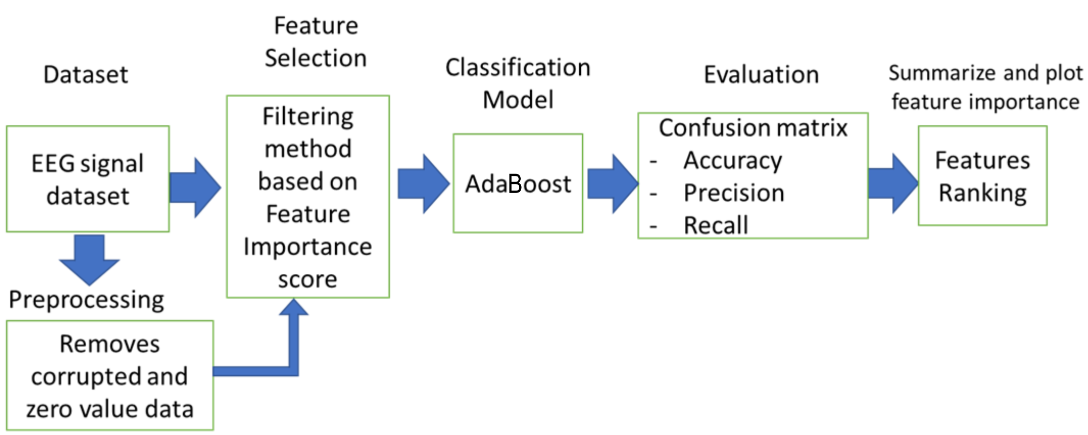

Figure 1. The feature selection of EEG signal based on feature importance score

Following are the stages of the feature selection and model training process using the selected features.

1) Provide the EEG dataset of 8 Indonesian words ("makan", "minum", "lapar", "haus", "senang", "sedih", "sakit" and "toilet").
2) There are two scenarios that were conducted, dataset applied without preprocessing and else with preprocessing.
3) Feature selection uses the filter method based on the feature importance scores.
4) The features are applied in AdaBoost model that has the advantage of resisting overfitting [43]. For all algorithm iterations, the samples set was fixed, and only its weights are changed [43]. The observation weights are initialized using

$$w_i = \frac{1}{N}, \tag{1}$$

where $i = 1, 2, \ldots, N$ for each training sample, where each sample belongs to the class $\{1, 2, \ldots, k\}$ [43].

5) The performance of the AdaBoost model is evaluated using confusion matrix, by calculating the predictive accuracy, precision and recall [44].

The accuracy, precision, and recall values are calculated using the following formulas [45]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

where TP (True positive): correctly classified; TN (True negative): correctly rejected; FP (False positive): incorrectly classified (type I error); FN (False negative): incorrectly rejected (type II error).

4. **Experimental Result.** In this section, we present the experimental result.

4.1. **Dataset.** The dataset used in this research is a primary dataset, obtained through the acquisition of EEG signals from eight Indonesian words, i.e., makan means eat (English), minum means drink (English), lapar means hungry (English), haus means thirsty (English), senang means happy (English), sedih means sad (English), sakit means sick (English) and toilet means toilet (English). Participants who carried out the acquisition were 11 people, 7 men and 4 women, with five different experiment paradigms or tasks, i.e., relax, look at picture that is related to eight predetermined words, read a word in a normal voice, read word in mind, imagining word with closed eyes. Each participant carried out 5 acquisitions, and data total 35200 samples and 70 features. Part of the dataset is shown in Table 1.

TABLE 1. Part of EEG dataset of eight words based on five acquisition schemes

SUBJECT	WORD	AF3_THETA	AF3_ALPHA	AF3_LOW_BETA	AF3_HIGH_BETA	AF3_GAMMA	F7_THETA	F7_ALPHA	F7_LOW_BETA
S01-1Aj	MAKAN	0.748956	1.743176	0.983523	1.96705	1.283365	0.544091	1.111515	1.014992
S01-1Aj	MAKAN	0.807715	2.006612	0.799029	2.017621	1.241508	0.605498	1.231996	0.855567
S01-1Aj	MINUM	1.25597	0.476775	1.362341	1.383737	1.008682	1.374154	0.909555	0.805395
S01-1Aj	MINUM	1.200756	0.366774	1.322671	1.417823	0.924297	1.453437	0.750904	0.660121
S01-1Aj	LAPAR	15.3334	2.138177	0.544348	1.444581	0.508819	2.13894	1.198907	0.67257
S01-1Aj	LAPAR	17.94966	2.251893	0.568215	1.520523	0.639559	2.389161	1.020033	0.718182
S01-1Aj	HAUS	2.234049	1.16	0.522923	0.961064	1.622324	1.258841	1.148527	0.391589
S01-1Aj	HAUS	2.62482	0.8075	0.457007	0.932137	1.707069	1.311642	0.855363	0.313063
S01-1Aj	SENANG	4.655643	1.869089	0.637621	3.112569	0.620939	4.111645	1.217647	0.569077
S01-1Aj	SENANG	4.468931	2.925513	0.753749	3.545907	0.597763	4.222561	1.954325	0.612835
S01-1Aj	SEDIH	12.16555	1.657921	0.556077	0.689249	0.342006	3.90517	0.742025	0.327777
S01-1Aj	SEDIH	5.973289	1.231698	0.504166	0.74919	0.295562	2.660531	0.771201	0.369256
S01-1Aj	SAKIT	3.18792	0.794446	0.773643	0.653096	0.449018	1.006206	0.528925	0.585605
S01-1Aj	SAKIT	1.992047	0.531172	0.636414	0.785457	0.447994	0.957407	0.563946	0.510291
S01-1Aj	TOILET	33.31791	3.464122	1.334854	1.485933	0.528216	24.96496	1.703305	0.480413
S01-1Aj	TOILET	22.05128	2.486174	1.321541	1.784595	0.549228	20.36251	1.533454	0.5185

4.2. **Preprocessing.** There are two scenarios that were conducted, dataset applied without preprocessing and else with preprocessing. When data are applied with preprocessing step, the raw dataset with 35200×70 was reduced by removing the missing value, zero value and labeled the output variable (class), and then the dataset is applied to feature importance score.

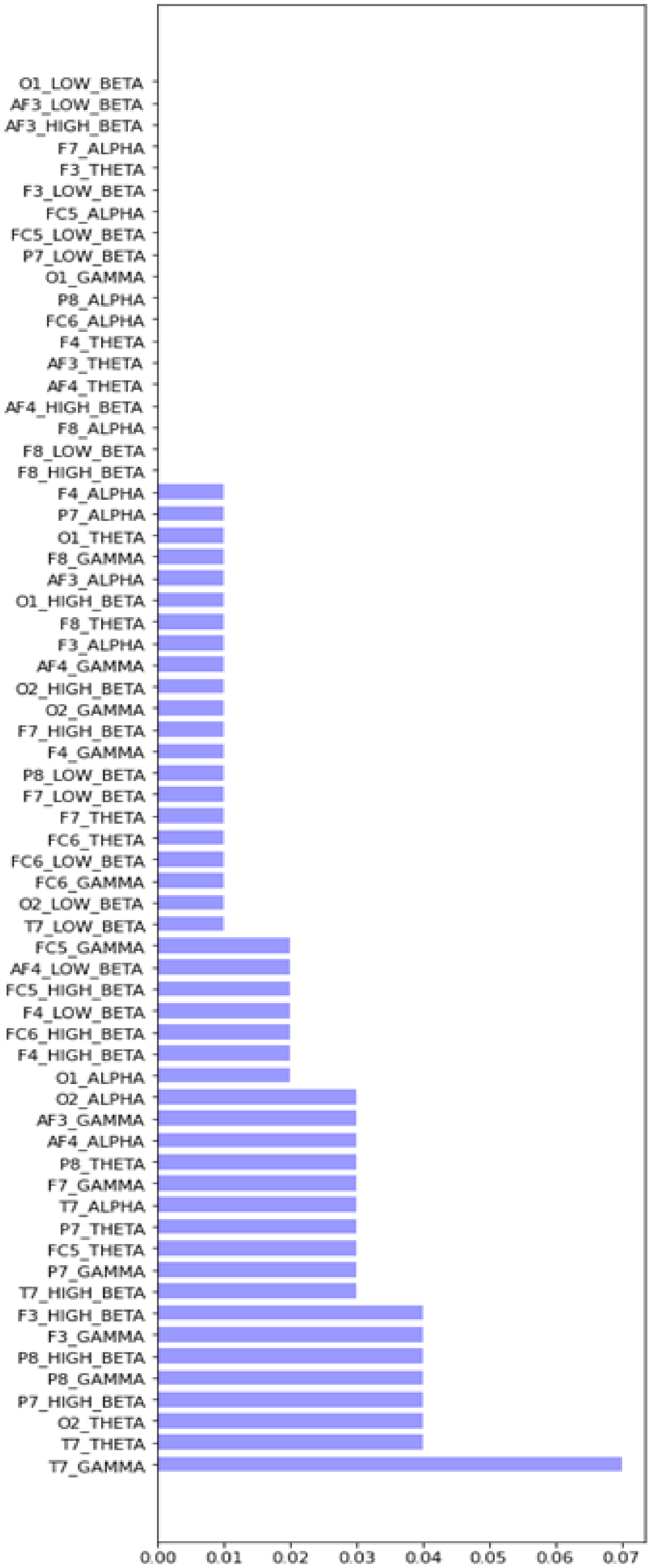

FIGURE 2. Feature ranking based on feature importance score

4.3. **Feature selection.** The feature filtering method was applied for feature selection by calculating the feature importance scores and the results are shown in Figure 2.

As shown in Figure 2, we obtained top ten features based on feature importance score are T7_GAMMA, T7_THETA, P7_HIGH BETA, P8_GAMMA, P8_HIGH BETA, F3_GAMMA, F3_HIGH BETA, T7_HIGH BETA, P7_GAMMA and FC5_THETA.

4.4. **Classification and evaluation.** The features that were obtained through feature importance score were then applied to AdaBoost model, for ten times and obtained the accuracy, precision and recall values as shown in Table 2.

TABLE 2. Performances of AdaBoost classifier model

Tes	Without preprocessing			With preprocessing		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
1	0.50	0.59	0.48	0.76	0.80	0.85
2	0.50	0.59	0.48	0.75	0.80	0.84
3	0.51	0.60	0.47	0.76	0.80	0.83
4	0.51	0.59	0.50	0.76	0.80	0.84
5	0.50	0.59	0.50	0.76	0.80	0.83
6	0.51	0.60	0.47	0.76	0.80	0.84
7	0.51	0.60	0.47	0.75	0.80	0.84
8	0.50	0.58	0.48	0.75	0.79	0.84
9	0.51	0.60	0.48	0.76	0.80	0.84
10	0.50	0.59	0.48	0.75	0.80	0.84
Mean	**0.50**	0.59	0.48	**0.75**	0.80	0.84

Table 2 presents the performance values of the AdaBoost model based on accuracy, precision, and recall parameters. The model was ten times running, for both the scenarios, and then the accuracy, precision, and recall were calculated. The average was calculated from ten experiments and obtained the accuracy value of the data that has been through the preprocessing is 25% higher, the precision value is 21% higher and the recall is 36% greater than the initial data.

5. **Conclusions.** Based on the experiment, we obtained the accuracy value of the data that has been through the preprocessing is 25% higher, the precision value is 21% higher and the recall is 36% greater than the initial data. For the feature selection process using the feature importance score, the best features of dataset in the AdaBoost model are T7_GAMMA, T7_THETA, P7_HIGH BETA, P8_GAMMA, P8_HIGH BETA, F3_GAMMA, F3_HIGH BETA, T7_HIGH BETA, P7_GAMMA and FC5_THETA. For the future we will analyze the features for another model and the brain region corellation.

REFERENCES

[1] M. M. Alsaleh, M. Arvaneh, H. Christensen and R. K. Moore, Brain-computer interface technology for speech recognition: A review, *2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA2016)*, 2016.
[2] M. Koctúrová and J. Juhár, An overview of BCI-based speech recognition methods, *Proc. of IEEE World Symposium on Digital Intelligence for Systems and Machines (DISA2018)*, pp.327-330, DOI: 10.1109/DISA.2018.8490536, 2018.
[3] C. Cooney, R. Folli and D. Coyle, Neurolinguistics research advancing development of a direct-speech brain-computer interface, *ISCIENCE*, vol.8, pp.103-125, 2018.
[4] N. Hashim, A. Ali and W. N. Mohd-Isa, Word-based classification of imagined speech using EEG, in *Computational Science and Technology. ICCST 2017. Lecture Notes in Electrical Engineering*, R. Alfred, H. Iida, A. A. Ibrahim and Y. Lim (eds.), Singapore, Springer, 2018.
[5] M. O. Tamm, Y. Muhammad and N. Muhammad, Classification of vowels from imagined speech with convolutional neural networks, *Computers*, vol.9, no.2, 2020.

[6] B. M. Idrees and O. Farooq, Vowel classification using wavelet decomposition during speech imagery, *The 3rd Int. Conf. Signal Process. Integr. Networks (SPIN2016)*, pp.636-640, 2016.
[7] B. M. Idrees and O. Farooq, EEG based vowel classification during speech imagery, *Proc. of the 3rd Int. Conf. Comput. Sustain. Glob. Dev. (INDIACom2016)*, pp.1130-1134, 2016.
[8] L. C. Sarmiento, J. B. Rodríguez, O. López, S. I. Villamizar, R. D. Guevara and C. J. Cortes-Rodriguez, Recognition of silent speech syllables for brain-computer interfaces, *2019 IEEE International Conference on E-Health Networking, Application & Services (HealthCom)*, 2019.
[9] S. Deng, R. Srinivasan, T. Lappas and M. D'Zmura, EEG classification of imagined syllable rhythm using Hilbert spectrum methods, *J. Neural Eng.*, vol.7, no.4, 2010.
[10] M. N. I. Qureshi, B. Min, H. Park, D. Cho, W. Choi and B. Lee, Multiclass classification of word imagination speech with hybrid connectivity features, *IEEE Trans. Biomed. Eng.*, vol.6, pp.2168-2177, 2017.
[11] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam and S. Moghimi, EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density, *Biomed. Signal Process. Control*, vol.59, 2020.
[12] B. Boloukian and F. Safi-Esfahani, Recognition of words from brain-generated signals of speech-impaired people: Application of autoencoders as a neural turing machine controller in deep neural networks, *Neural Networks*, vol.121, pp.186-207, 2020.
[13] D. Pawar and S. Dhage, Multiclass covert speech classification using extreme learning machine, *Biomed. Eng. Lett.*, vol.10, no.2, pp.217-226, 2020.
[14] M. AlSaleh, R. Moore, H. Christensen and M. Arvaneh, Discriminating between imagined speech and non-speech tasks using EEG, *2018 the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp.1952-1955, 2018.
[15] M. Alsaleh, R. Moore, H. Christensen and M. Arvaneh, Examining temporal variations in recognizing unspoken words using EEG signals, *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp.976-981, 2018.
[16] A. R. Sereshkeh, R. Yousefi, A. T. Wong and T. Chau, Online classification of imagined speech using functional near-infrared spectroscopy signals, *J. Neural Eng.*, vol.16, no.1, pp.1-13, 2019.
[17] N. Abdallah, B. Daya, S. Khawandi and P. Chauvet, Electroencephalographic based brain computer interface for unspoken speech, *2017 Sensors Networks Smart Emerg. Technol. (SENSET2017)*, pp.1-4, 2017.
[18] X. Liu, J. Shen and W. Zhao, Epileptic EEG identification based on hybrid feature extraction, *J. Mech. Med. Biol.*, vol.20, no.6, 2020.
[19] J. Brownlee, *Data Preparation for Machine Learning*, https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/, 2019.
[20] S. V. Eslahi, N. J. Dabanloo and K. Maghooli, A GA-based feature selection of the EEG signals by classification evaluation: Application in BCI systems, *arXiv.org*, arXiv: 1903.02081, 2019.
[21] J. Brownlee, *Machine Learning Performance Improvement Cheat Sheet*, https://machinelearning mastery.com/machine-learning-performance-improvement-cheat-sheet/, 2016.
[22] L. Duan, M. Bao, S. Cui, Y. Qiao and J. Miao, Motor imagery EEG classification based on kernel hierarchical extreme learning machine, *Cognitive Computation*, vol.9, pp.758-765, 2017.
[23] A. Liu, K. Chen, Q. Liu, Q. Ai, Y. Xie and A. Chen, Feature selection for motor imagery EEG classification based on firefly algorithm and learning automata, *Sensors (Switzerland)*, vol.17, no.11, 2017.
[24] H.-H. Huang, A. Condor and H. J. Huang, Classification of EEG motion artifact signals using spatial ICA, in *Statistical Modeling in Biomedical Research. Emerging Topics in Statistics and Biostatistics*, Y. Zhao and D. G. Chen (eds.), Cham, Springer, 2020.
[25] F. Tang, L. Adam and B. Si, Group feature selection with multiclass support vector machine, *Neurocomputing*, vol.317, pp.42-49, 2018.
[26] M. A. Asghar et al., EEG-based multi-modal emotion recognition using bag of deep features: An optimal feature selection approach, *Sensors (Switzerland)*, vol.19, no.23, 2019.
[27] A. C. Ramos and M. Vellasco, Quantum-inspired evolutionary algorithm for feature selection in motor imagery EEG classification, *Proc. of 2018 IEEE Congr. Evol. Comput. (CEC2018)*, pp.1-8, 2018.
[28] U. M. Khaire and R. Dhanalakshmi, Stability of feature selection algorithm: A review, *J. King Saud Univ. – Comput. Inf. Sci.*, 2019.
[29] M. H. Bhatti et al., Soft computing-based EEG classification by optimal feature selection and neural networks, *IEEE Trans. Ind. Informatics*, vol.15, no.10, pp.5747-5754, 2019.
[30] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 2013.

[31] J. Ma, B. Xue and M. Zhang, A hybrid filter-wrapper feature selection approach for authorship attribution, *International Journal of Innovative Computing, Information and Control*, vol.15, no.5, pp.1989-2006, 2019.

[32] S. Rumao, *Exploration of Variable Importance and Variable Selection Techniques in Presence of Correlated Variables*, Master Thesis, Department of Mathematical Sciences, College of Science, Rochester Institute of Technology, 2019.

[33] F. Yang, X. Zhao, W. Jiang, P. Gao and G. Liu, Multi-method fusion of cross-subject emotion recognition based on high-dimensional EEG features, *Front. Comput. Neurosci.*, vol.13, no.8, 2019.

[34] M. A. Rahman, F. Khanam, M. Ahmad and M. S. Uddin, Multiclass EEG signal classification utilizing Rényi min-entropy-based feature selection from wavelet packet transformation, *Brain Informatics*, vol.7, no.1, 2020.

[35] A. Tiwari and A. Chaturvedi, A multiclass EEG signal classification model using spatial feature extraction and XGBoost algorithm, *IEEE Int. Conf. Intell. Robot. Syst.*, pp.4169-4175, 2019.

[36] J. H. Hur, S. Y. Ihm and Y. H. Park, A variable impacts measurement in random forest for mobile cloud computing, *Wirel. Commun. Mob. Comput.*, 2017.

[37] M. Z. Baig, N. Aslam and H. P. H. Shum, Filtering techniques for channel selection in motor imagery EEG applications: A survey, *Artif. Intell. Rev.*, vol.53, no.2, pp.1207-1232, 2020.

[38] M. P. Ellies-Oury, M. Chavent, A. Conanec, M. Bonnet, B. Picard and J. Saracco, Statistical model choice including variable selection based on variable importance: A relevant way for biomarkers selection to predict meat tenderness, *Sci. Rep.*, vol.9, no.1, pp.1-12, 2019.

[39] M. J. Hallett, J. J. Fan, X. G. Su, R. A. Levine and M. E. Nunn, Random forest and variable importance rankings for correlated survival data, with applications to tooth loss, *Stat. Modelling*, vol.14, no.6, pp.523-547, 2014.

[40] P. Wei, Z. Lu and J. Song, Variable importance analysis: A comprehensive review, *Reliab. Eng. Syst. Saf.*, vol.142, no.3, pp.399-432, 2015.

[41] I. A. Dfuf, J. F. Perez-Minayo, J. M. M. McWilliams and C. G. Fernandez, Variable importance analysis in imbalanced datasets: A new approach, *IEEE Access*, vol.8, pp.127404-127430, 2020.

[42] G. Badih, M. Pierre and B. Laurent, Assessing variable importance in clustering: A new method based on unsupervised binary decision trees, *Comput. Stat.*, vol.34, no.1, pp.301-321, 2019.

[43] R. Bose, S. K. Goh, K. F. Wong, N. Thakor, A. Bezerianos and J. Li, Classification of brain signal (EEG) induced by shape-analogous letter perception, *Adv. Eng. Informatics*, vol.42, no.9, 2019.

[44] R. Spencer, F. Thabtah, N. Abdelhamid and M. Thompson, Exploring feature selection and classification methods for predicting heart disease, *Digit. Heal.*, vol.6, 2020.

[45] A. Kulkarni, D. Chong and F. A. Batarseh, *Foundations of Data Imbalance and Solutions for a Data Democracy*, Elsevier Inc., 2020.