Turkish Journal of Science & Technology
17(1), 135-141, 2022

Research Paper
https://doi.org/10.55525/tjst.1074540

A new dataset for EEG abnormality detection: MTOUH

İrem TAŞCI[1], Burak TAŞCI[2*], Şengül DOĞAN[3], Türker TUNCER[4]
[1] Department of Neurology, School of Medicine, Malatya Turgut Ozal University, Malatya, Turkey
[2] Vocational School of Technical Sciences, Firat University, Elazig 23119, Turkey
[3,4]Department of Digital Forensics Engineering, Technology Faculty, Firat University, Elazig, Turkey
[1]irem.tasci@ozal.edu.tr, *[2] btasci@firat.edu.tr, [3] sdogan@firat.edu.tr, [4] turkertuncer @firat.edu.tr

(Geliş/Received: 16/02/2022; Kabul/Accepted: 06/03/2022)

Abstract: Electroencephalogram (EEG) is widely used for monitoring electrical activity in the brain. Analyzing EEG signals by physicians is tiring and time-consuming. Therefore, machine learning techniques can be used to improve detection accuracy. This study created a 2-class data set consisting of 35 channels, 10575x15 seconds of normal and 11240x15 seconds of abnormal EEG signals. This data set was obtained by examining the EEG signals of the patients who applied to Malatya Turgut Ozal University Malatya Research and Training Hospital in 2021. In the study, a statistical feature extraction-based model is proposed. After the feature vector reduction was performed using the neighboring component analysis to the proposed model, the classification was made using the support vector machines. The highest accuracy out of 35 channels was obtained in the P4O2 channel. Accuracy, sensitivity, specificity, precision and f-score for the P4O2 channel were 81.3%, 78.9%, 83.7%, 82.0% and 80.4%, respectively.

Key words: Electroencephalography signals classification, SVM, Machine Learning.

EEG anormallik tespiti için yeni bir veri seti: MTOUH

Öz: Elektroensefalogram (EEG), beyindeki elektriksel aktivitenin izlenmesi için yaygın olarak kullanılmaktadır. EEG sinyallerinin hekimler tarafından incelenmesi yorucu ve zaman alıcıdır. Bu nedenle, algılama doğruluğunu artırmak için makine öğrenme teknikleri kullanılabilir. Bu çalışmada 35 kanal, 10575x15 saniye normal ve 11240x15 saniye anormal EEG sinyalinden oluşan 2 sınıflı veri seti oluşturulmuştur. Bu veri seti Turgut Özal Üniversitesi Malatya Eğitim Araştırma Hastanesi' ne 2021 yılında başvuran hastaların EEG sinyalleri incelenerek elde edilmiştir. Çalışmada istatistiksel özellik çıkarımı tabanlı bir model önerilmiştir. Önerilen modele komşu bileşen analizi kullanılarak öznitelik vektörü indirgemesi yapıldıktan sonra destek vektör makineleri kullanılarak sınıflandırma yapılmıştır. 35 kanaldan en yüksek doğruluk P4O2 kanalında elde edilmiştir. P4O2 kanalı için doğruluk, duyarlılık, özgüllük, kesinlik ve f-skoru sırasıyla %81.3,%78.9, %83.7, %82.0 ve %80.4 olarak elde edilmiştir.

Anahtar kelimeler: Elektroensefalografi sinyal sınıflandırması, DVM, Makine Öğrenmesi.

1. Introduction

Electroencephalography (EEG) is the process of recording the brain's electrical activity through electrodes placed on the scalp [1]. The German neuropsychiatrist Hans Berger used the term EEG and performed the first human EEG recording [2]. EEG has been utilized in the diagnosis and differential diagnosis of various neurological illnesses since the turn of the twentieth century. It has been used to diagnose epilepsy, non-epileptic psychogenic seizures, hypoxia, and intracranial space-occupying lesions [3]. Postsynaptic electrical potentials of pyramidal neurons in the cortex play a role in the formation of EEG activity. Five different brain waves have been defined according to their frequency ranges. Delta rhythm is 0.5-3.5 Hz, theta rhythm is 4-7.5 Hz, the alpha rhythm is 8-14 Hz, the beta rhythm is 15-30 Hz, gamma rhythm is 30-48 Hz. In a healthy person, the dominant rhythm is alpha in the parietooccipital regions with eyes closed and at rest. Beta rhythm is observed while awake and eyes open. Theta is the dominant rhythm in shallow sleep and delta in deep sleep and anesthesia [4]. The electrodes are attached to the scalp during EEG recording according to the 10-20 system accepted by the international federation. Recordings can be performed with unipolar or bipolar montage. The most commonly used is the bipolar mount. In this technique, electrodes are attached to the right and left prefrontal (Fp), frontal (F), central (C), temporal (T), parietal (P), occipital (O), auricular (A) regions. Odd-numbered electrodes indicate the left hemisphere, and even-numbered electrodes indicate right hemisphere localizations [5]. There are many clinical and engineering studies on the use of EEG signals to detect neurological diseases. Some studies on EEG signals in the literature are presented below.

[*] Corresponding author: btasci@firat.edu.tr. ORCID Number of authors: [1] 0000-0001-7069-769X, *[2] 0000-0002-4490-0946, [3]0000-0001-9677-5684, [4]0000-0002-1425-4664

Zhao et al. [6] classified the EEG signals using the 1D CNN deep learning structure. The accuracy rates obtained in their study were 99.52% in the two-class classification problem, 98.06% in the three-class EEG classification problem, and 93.55% in the five-class classification problem, respectively. Khan et al. [7] classified the features selected by correlation-based Q-score using an LSTM-based deep learning model after extracting the features of the EEG signal using HVD (Hilbert Vibration Decomposition). Their studies were conducted with two different data sets. An accuracy of 96.00% was achieved for the Bonn dataset and 83.30% for the Sensor Networks Research Lab data. Wang et al. [8] performed feature extraction from EEG signals using EMD and DWT (Discrete Wavelet Transform) methods. Their study obtained 92.07% accuracy, 91.13% sensitivity, and 92.96% specificity for the Bonn data set. Rashid et al. [9] obtained accuracy values (99.21%, 93.19%, 93.57%, and 90.32%) for CI Competition III, IVA, and BCI Competition IV datasets, respectively, using the kNN technique. Kumar et al. [10] transformed the real value mode into an analytical signal with frequency spectrum by VMD (Variable Mode Decomposition) method. Then, semantic feature extraction was applied to generate the features. A Random Forest classifier was used in the study, and a success rate of 94.1% was achieved. Sheoran et al. [11] obtained scalogram images by transitioning EEG signals from the time domain to frequency domain with CWT (Continuous Wavelet Transform). Feature extraction was performed by calculating the potential feature values of the instantaneous frequency components, LBP (Local Binary Patterns), and HOG (Oriented Gradient Histograms) from the obtained scalogram images. They obtained an accuracy value of 99.08% with the SVM classifier. Bera et al. [12] achieved a success rate of 98% for binary class and 84% for multiclass classification with the error correction exit codes (ECOC) approach using the BCI Competition-IV dataset. Ha et al [13] firstly, the motor image EEG signals in the BCI Competition-IV dataset were converted into 2D images using the short-term Fourier transform (STFT) algorithm. The converted signals were then used for training and testing the capsule network. In this study, 78.44% accuracy was obtained.

2. Material and Method

2.1. Dataset

Ethical approval of the study was given by the Malatya Turgut Ozal University Medical Faculty Ethics Committee (2022/01), following the principles of the Declaration of Helsinki. Awake EEG images of patients over 18 who applied to the electroneurophysiology laboratory were scanned retrospectively. EEG recordings were made in the awake state with bipolar mounting according to the 10-20 system accepted by the International Federation. In this technique, electrodes are attached to the right and left prefrontal (Fp), frontal (F), central (C), temporal (T), parietal (P), occipital (O), auricular (A) regions. Electrodes indicated with odd numbers indicate their localizations in the left hemisphere, and even numbers indicate their localizations in the right hemisphere. EEG signal recordings of the patients were recorded in 500Hz EDF format. These signals were labeled as normal and abnormal after being reconstructed as 15-second data packets with the help of the Matlab program for 35 channels. Patients under 18, patients who had sleep EEG recordings, and those with intense artifacts were excluded from the study. Patient information and details are given in Table 1.

Table 1. Dataset information

Classes	**Male**	**Female**	**Age**	**Number of Channels**	**Number of Data Packs**
Normal	20	24	32.4±11.28	35	10575x15sec
Abnormal	113	127	38.4±18.8	35	11240x15sec

Furthermore, this database was published publicly and the users/researcher can download this database from https://www.kaggle.com/buraktaci/mtouh (accessed on 3 March 2022).

2.2. Method

In this work, a statistical feature extraction-based model has been presented. The presented model has been applied to the Collected EEG signal dataset. This model contains three primary phases: statistical feature extraction, feature vector reduction using neighborhood component analysis, and classification with support vector machine classifier. To better explain the presented model, a graph of this model is given in Figure 1.

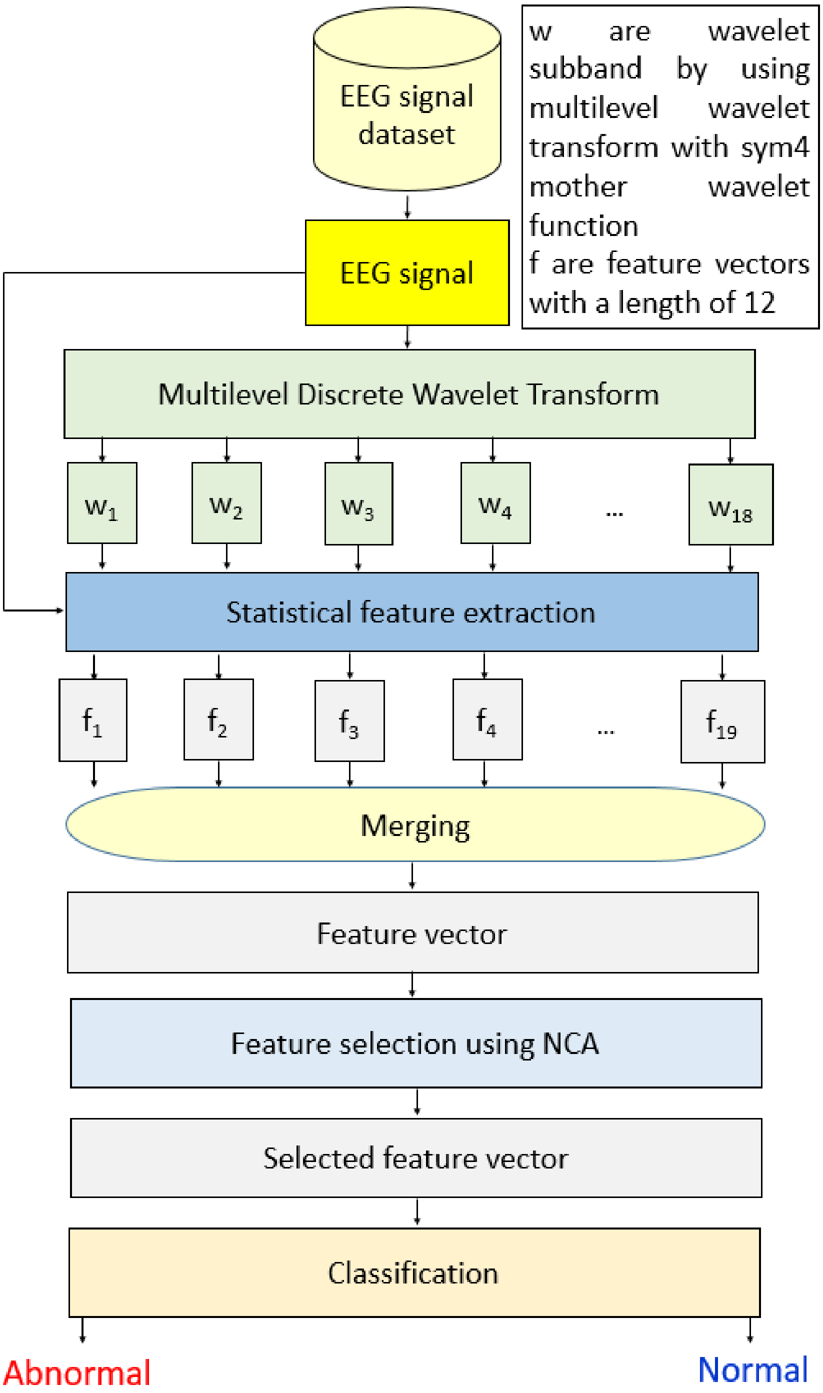
EEG signal dataset
w are wavelet subband by using multilevel wavelet transform with sym4 mother wavelet function
f are feature vectors with a length of 12
EEG signal
Multilevel Discrete Wavelet Transform
w_1
w_2
w_3
w_4
...
w_{18}
Statistical feature extraction
f_1
f_2
f_3
f_4
...
f_{19}
Merging
Feature vector
Feature selection using NCA
Selected feature vector
Classification
Abnormal
Normal

Figure 1. Graph of the presented EEG signal classification model using a statistical model.

The steps of the presented model are given below.

Step 0: Read each EEG signal from the dataset.

Step 1: Apply multilevel DWT to EEG signal to generate subbands.

$$[low^1, high^1] = DWT(signal,' sym4') \quad (1)$$

$$[low^{t+1}, high^{t+1}] = DWT(low^t,' sym4'), t \in \{1,2, \dots ,8\} \quad (2)$$

$$w^{2h-1} = low^{h}, h \in \{1,2,\dots,9\} \tag{3}$$

$$w^{2h} = high^{h} \tag{4}$$

Herein, low and $high$ denote low-pass and high-pass filter subbands of the DWT ($DWT(.,.)$), w contains wavelet subbands. We generated 18 wavelet subbands using multilevel DWT, and both low and high subbands have been utilized to extract features in this work.
Step 2: Generated statistical features from EEG signal and wavelet subbands (w) to create a feature vector.

$$f^{1} = sfg(signal) \tag{5}$$

$$f^{t} = sfg(w^{t-1}), t \in \{2,3,\dots,19\} \tag{6}$$

Herein, f are generated feature vectors, and the length of each feature vector is 12. To extract statistical features, maximum, minimum, average, median, mode, standard deviation, information entropy, root mean square error, range, mean absolute deviation, variance, skewness, and kurtosis moments have been used. By applying these moments, 12 statistical features have been extracted from each generated one-dimensional signal. Briefly, sfg(.) extracts 12 features from a signal. In this step, 19 feature vectors have been generated.
Step 3: Merge the extracted feature vectors to obtain the final feature vector.

$$fv\big(i + 12 \times (j-1)\big) = f^{j}(i), i \in \{1,2,\dots,12\}, j \in \{1,2,\dots,19\} \tag{7}$$

Herein, fv is a feature vector with a length of 228 (=12×19).
Step 3: Normalize fv applying minimum-maximum normalization.
Step 4: Choose the most informative/distinctive 60 features employing NCA [14] feature selection model. NCA is the feature selection version of the k nearest neighbors (kNN) [15] method and uses distances to assign weights to each feature. We have selected 60 of 228 features.
Step 5: Classify the chosen feature using the Fine Gaussian Support Vector Machine (FG-SVM) [16] [17] classifier with an 80:20 split ratio.

3. Performance Analysis

3.1. Experimental setup and Results

All coding in this study was carried out with a simulation program called Matlab, which is based on the Windows 10 operating system. 80% of the data set used to evaluate the performance of the proposed method is randomly allocated for training and 20% for testing.
Precision, sensitivity, specificity, F1-score, and accuracy metrics were used to obtain results [18, 19]. These metrics are generally used in machine learning. Therefore, we considered these metrics to evaluate our proposed method. The mathematical definition of these performance metrics is given in Equations 8-12. Also, true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) values are used to calculate these performance metrics. In the study, classification was made with Fine Gaussian SVM. Classification has been performed for 35 different channels. The confusion matrices of the first 6 channels with the best results are given in Figure 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$F - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{12}$$

Table 2. Channel-based results of the proposed method (%)

Channel	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F-Score(%)
P4O2	**81.3**	**78.9**	83.7	82.0	**80.4**
C3P3	81.1	74.7	**87.1**	**84.5**	79.3
P3O1	80.9	78.3	83.3	81.5	79.9
C4P4	80.3	76.6	83.8	81.7	79.1
F3C3	79.7	73.1	85.8	82.9	77.7
F4C4	79.4	76.6	82.1	80.1	78.3
P3A1	79.3	75.9	82.4	80.2	78.0
C4A2	78.8	75.0	82.3	79.9	77.4
T6A2	78.7	70.5	86.4	83.0	76.3
T4T6	78.4	70.7	85.8	82.6	76.2
F7A1	78.2	75.2	81.0	78.8	77.0
F8T4	78.1	71.8	84.0	80.9	76.1
F3A1	78.1	73.8	82.2	79.6	76.5
T3T5	78.0	71.8	83.9	80.8	76.0
T6O2	77.6	70.2	84.5	81.0	75.2
C3A1	77.5	73.2	81.5	78.9	75.9
O2A2	77.1	70.9	82.8	79.5	75.0
F4A2	76.8	74.0	79.4	77.1	75.5
P1F3	76.6	70.5	82.2	78.9	74.5
O1A1	76.4	72.4	80.1	77.4	74.8
P4A2	76.4	72.2	80.2	77.5	74.8
T5O1	76.1	72.8	79.3	76.8	74.7
P1F7	75.8	71.4	80.0	77.0	74.1
P1A1	75.6	67.8	82.9	78.9	72.9
T4A2	75.5	66.1	84.3	79.9	72.4
F7A1	75.5	68.1	82.4	78.4	72.9
CZA1	75.2	65.6	84.3	79.7	72.0
T5A1	75.1	67.5	82.3	78.2	72.5
P2F4	75.0	66.1	83.3	78.8	71.9
P2F8	75.0	66.1	83.3	78.8	71.9
PZA2	74.9	74.6	75.3	73.9	74.2
P2A2	74.9	66.8	82.5	78.2	72.1
F8A2	74.7	66.0	82.8	78.3	71.6
T3A1	74.4	68.6	79.9	76.2	72.2
FZA2	74.2	67.4	80.6	76.6	71.7

The highest results for accuracy sensitivity and F-Score, with 78.9% and 80.4%, respectively, were obtained at the P4O2 channel. The highest results for specificity and precision, 87.1% and 84.5%, respectively, were obtained in the C3P3 channel. Lowest accuracy at 74.2% at FZA2 channel, lowest sensitivity at 66.0% at F8A2 channel, lowest specificity at 79.3% at T5O1 channel, lowest precision at 73.9% at PZA2 channel, and lowest F-score(%) at FZA2 channel at 71.7% has been obtained.

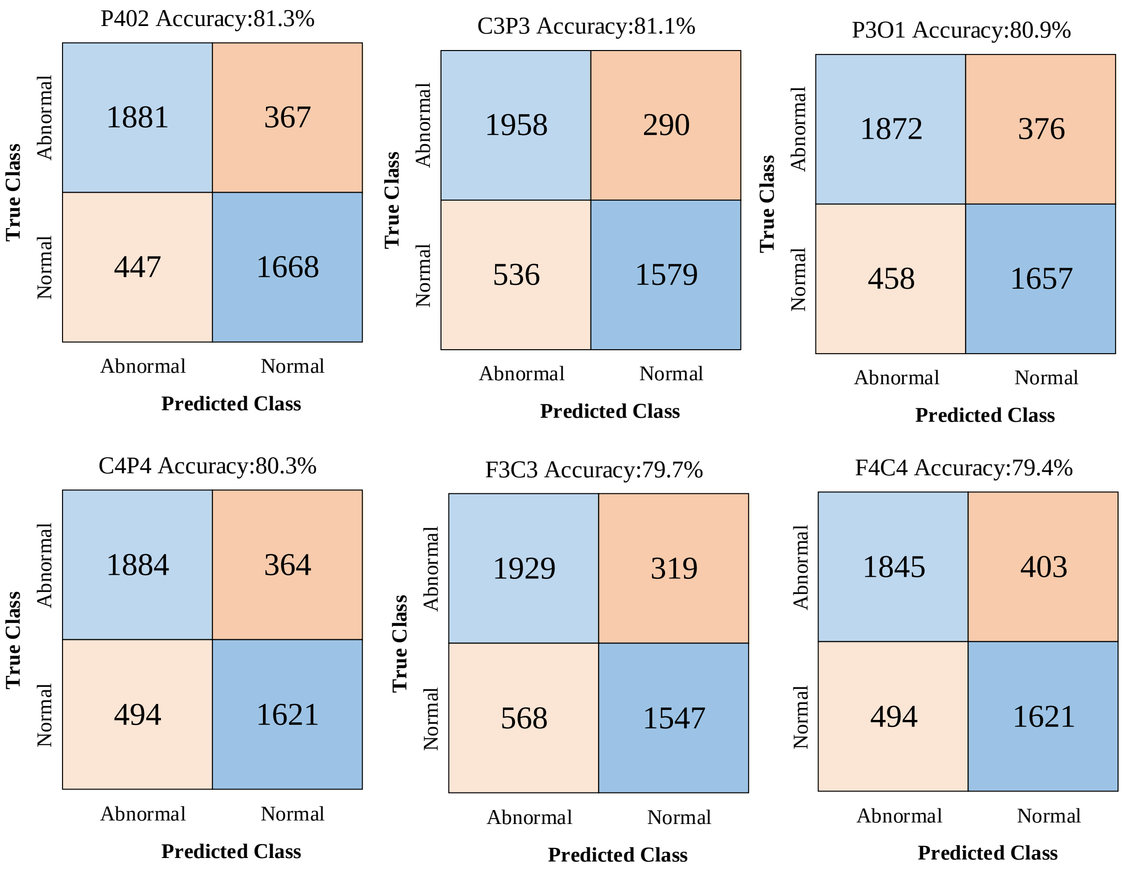

Figure 2. Confusion matrices of 6 channels.

Performance metrics are calculated using these confusion matrices. The performance ratios calculated for 35 channels are tabulated in Table 2.

4. Conclusions

In this study, a feature selection-based decision support system was proposed to detect abnormal EEG signals. In the study, feature extraction was performed for each channel difference of the EEG. The highest classification accuracy was obtained with the P4O2 channel. In the process, it is aimed to increase the classification performance and accuracy and reduce the cost. It is expected that the study will help physicians in diagnosis.

The primary purpose of this article is to classify abnormal EEG signals in the newly created data set. In future studies, the number of data and classes in the dataset will be increased.

References

[1] Kocaaslan A., Bayazıt O., Kahya M., Elektroensefalografinin Biyofiziksel Temelleri, Turkiye Klinikleri J Neurol 2017: 10(2); 110-114.

[2] Biasiucci A., Franceschiello B., Murray M.M., Electroencephalography, Current Biology 2019: 29; 80-85.

[3] Millett D., Hans Berger: From psychic energy to the EEG, Perspectives in biology and medicine 2001: 44; 522-542.

[4] Galip A., Sabiha T., Elektroensefalografinin Tarihçesi, Turkiye Klinikleri J Neurol 2017: 10(2); 105-109.

[5] Süleyman K., Nihat Ş., Rutin Elektroensefalografi Kayıtlaması ve Aktivasyon Yöntemleri, Turkiye Klinikleri J Neurol, 2017: 10(2); 115-119.

[6] Zhao W., Zhao W., Wang W., Jiang X., Zhang X., Peng Y., Zhang B., Zhang G., A novel deep neural network for robust detection of seizures using EEG signals, Computational and mathematical methods in medicine: 2020;1-9.

[7] Khan P., Khan Y., Kumar S., Khan M.S., Gandomi A.H., HVD-LSTM based recognition of epileptic seizures and normal human activity, Computers in Biology and Medicine 2021: 136; 104684.

[8] Wang Y., Dai Y., Liu Z., Guo J., Cao G., Ouyang M., Liu D., Shan Y., Kang G., Zhao G., Computer-Aided Intracranial EEG Signal Identification Method Based on a Multi-Branch Deep Learning Fusion Model and Clinical Validation, Brain Sciences 2021; 11; 615.

[9] Rashid M., Bari B.S., Hasan M.J., Razman M.A.M., Musa R.M., Ab Nasir A.F., Majeed A.P.A., The classification of motor imagery response: an accuracy enhancement through the ensemble of random subspace k-NN, PeerJ Computer Science 2021; 7; e374.

[10] Ravi Kumar M., Srinivasa Rao Y., Epileptic seizures classification in EEG signal based on semantic features and variational mode decomposition, Cluster Computing 2019: 22; 13521-13531.

[11] Sheoran P., Rathee N., Saini J., Epileptic seizure detection using bidimensional empirical mode decomposition and distance metric learning on scalogram, in: 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2020, pp. 675-680.

[12] Bera S., Roy R., Sikdar D., Kar A., Mukhopadhyay R., Mahadevappal M., A randomised ensemble learning approach for multiclass motor imagery classification using error correcting output coding, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018; 5081-5084.

[13] Ha K.-W., Jeong J.-W., Motor imagery EEG classification using capsule networks, Sensors, 2019: 19; 2854.

[14] Goldberger J., Roweis S., Hinton G., Salakhutdinov R., Neighbourhood components analysis, in: Proceedings of the 17th International Conference on Neural Information Processing Systems, MIT Press, Vancouver, British Columbia, Canada, 2004: 513–520.

[15] Peterson L.E., K-nearest neighbor, Scholarpedia 2009: 4; 1883.

[16] Vapnik V., The Support Vector Method of Function Estimation, in: J.A.K. Suykens, J. Vandewalle (Eds.) Nonlinear Modeling: Advanced Black-Box Techniques, Springer US, Boston, MA 1998: 55-85.

[17] Vapnik V., The nature of statistical learning theory, Springer science & business media, 1999.

[18] Warrens M.J., On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index, Journal of classification 2008: 25; 177-183.

[19] Chicco D., Jurman G., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC genomics 2020: 21; 6.