TechRxiv
Powered by IEEE

Neuro-Tech Revolution: AI-EEG Integration for Advanced Depression Diagnosis

This paper was downloaded from TechRxiv (https://www.techrxiv.org).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

06-07-2023 / 10-07-2023

CITATION

Gabdrakhimov, Bekarys; Detlefsen, Nicki Skafte; Uyanik, Cihan; Ejaz, Osama; Khan, Muhammed Ahmed; Hasan, Muhammad Abul; et al. (2023). Neuro-Tech Revolution: AI-EEG Integration for Advanced Depression Diagnosis. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23633841.v1

DOI

10.36227/techrxiv.23633841.v1

Neuro-Tech Revolution: AI-EEG Integration for Efficient Depression Diagnosis

Bekarys Gabdrakhimov, *Member, IEEE*, Nicki Detlefsen, Cihan Uyanik, Osama Ejaz, Muhammed Ahmed Khan, Muhammad Abul Hasan, Saad Ahmed Qazi and Sadasivan Puthusserypady, *Senior Member, IEEE*

***Abstract*— Major depressive disorder (MDD) is a common mental disorder affecting the lives of about 280 million people and increasing rates of suicidal mortality. The current methods of diagnosis of depression are subjective, time-consuming, expensive, and inaccurate because of its heterogeneous symptoms that overlap with other disorders. In this paper, we exploit the potential of the fusion of artificial intelligence (AI) and electroencephalogram (EEG) to revolutionize the automatic diagnosis of depression and compare the classification performance of machine learning (ML) and deep learning (DL) based techniques. Results from the analysis of data recorded from 46 subjects (23 MDD and 23 Control) show that the ML methods, particularly the ensemble model with the Dempster-Shafer combination rule outperforms other models, achieving an accuracy of 99.62% and showing robustness to the variations in the data. Our work also includes a study on the effect of various hyper-parameters, in particular the number of EEG channels, feature selection methods, number of selected features, and segmentation length on the model performance. The AI-EEG integration can enhance the accuracy of diagnosis, enable personalized treatment plans, and improve patient outcomes. Continued research, development, and validation of AI algorithms, in conjunction with ethical considerations, will be crucial to harness the full potential of this technology in mental healthcare.**

***Index Terms*— Major Depressive Disorder (MDD), Electroencephalogram (EEG), Convolutional Neural Networks (CNN), Deep Learning (DL), Machine Learning (ML).**

I. Introduction

MAJOR depressive disorder (MDD), also known as clinical or unipolar depression, is a prevalent mental disorder affecting the lives of people of all genders and ages. According to the WHO, approximately 280 million people worldwide are suffering from depression, and it is responsible for almost three fourth of a million deaths annually [1] [2]. Such repercussions could be avoided if depression had been diagnosed and treated in its early stages.

MDD is a complex mental health disorder that can be challenging to diagnose accurately. Currently, its diagnosis is based on psychiatric interviews. The most popular diagnosing methods are the 10th revision of the International Classification of Diseases (ICD-10) developed by the WHO [3] and the 5th revision of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) developed by the American Psychiatric Association (APA) [4]. However, these methods are subjective, and the effectiveness varies depending on the cooperation of the subject and the doctor. Also, depression symptoms are complex and vary widely between individuals, which impedes its diagnosis. According to Østergaard et al., more than 1400 possible combinations of symptoms can result in the diagnosis of MDD, and also, the symptoms overlap with other similar disorders and syndromes [5]. Due to all of these complications, around 50% of all depressed subjects remain untreated [6].

To make the diagnosis of depression low-cost, effective, objective, and reliable, researchers have been looking into physiological data to discover biomarkers of MDD. There are various physiological measurement tools, such as functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and positron emission tomography (PET), that have been used to diagnose depression [7]–[9]. Out of these tools, EEG has clear advantages as it is easy to administer, is tolerated well, and is relatively low cost.

EEG measures the electrical activity of the brain through electrodes placed on the scalp. It provides valuable information about the brain's functioning and can detect abnormalities in neural activity. Traditionally, EEG has been used to identify specific brainwave patterns associated with various mental states and disorders. The intersection of artificial intelligence (AI) and EEG holds significant potential to revolutionize the diagnosis and treatment of depression.

AI algorithms can process and analyze large amounts of EEG data, identifying patterns and relationships that may not be immediately apparent to human observers. Machine learning (ML) techniques can be employed to train AI models to recognize specific EEG patterns associated with depression, enabling them to differentiate between healthy individuals and those with depressive symptoms.

Several methods have been suggested to diagnose depression, and studies report that MDD and healthy subjects have differences in their EEG activity [10] [11]. These differences include amplitude, entropy, power in frequency bands, as well

B. Gabdrakhimov, C. Uyanik and S. Puthusserypady are with the Department of Health Technology, Technical University of Denmark, Kgs. Lyngby 2800, Denmark (e-mail: s210127@student.dtu.dk, {ciuya, sapu}@dtu.dk).

Nicki S. Detlefsen is with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby 2800, Denmark (e-mail: nsde@dtu.dk).

Muhammad A. Khan is with the Department of Electrical Engineering, Stanford University, Palo Alto, 94304, USA (e-mail: muhkhan@stanford.edu).

Muhammad A. Hassan, Saad A. Qazi and Osama Ejaz are with the Neurocomputation Lab, NED University of Engineering and Technology, Karachi, 75600, Pakistan (e-mail: abulhasan@neduet.edu.pk, saadqazi@neduet.edu.pk, osamaejaz@neduet.edu.pk).

Manuscript received xxx xx, 2023; revised xxx xx xxxx.

as differences between left and right hemispheres. Also, with the advent of ML and deep learning (DL) techniques, it has been possible to differentiate depressed and healthy subjects accurately [12] [13] [14]. However, for most of the works, the trained model is tested on the same dataset. There are many other variables that can affect the model performance, such as the sampling frequency, number of electrodes, montage used, etc. Therefore, to make the models generalizable and accessible, they need to perform well on other (unseen) datasets. Also, there are various hyper-parameters including EEG channels selected, segmentation length, feature selection method, and number of selected features, to be optimized. The effect of these parameters on the model performance should be tested to propose a standard way of designing depression diagnosis methods. In addition, the accuracy that can be achieved by ML and DL models vary significantly, and it is not obvious which method should be opted for. ML and DL methods will be compared not only in terms of accuracy but also in terms of generalization, speed and complexity.

This paper is organized as follows: A description of the datasets and proposed methodology is provided in Section II. Section III presents DL model architecture with the theoretical background of each layer. Section IV presents the results for ML and DL models. The interpretation and discussion of the obtained results are also been provided in this section. Section V concludes the paper with some suggestions for future research.

II. Methods

A. Data description

1) Dataset 1 (D1): The resting-state EEG data was obtained from 46 participants (23 healthy and 23 depressive). All participants were between the ages of 20 and 35, and none have a self-reported neurological disorder history. EEG data was acquired for 2 minutes each for eyes open (EO) and eyes closed (EC) states with 1-minute breaks in between. Mitsar NVX-52 EEG acquisition system was used with a 31 channel configuration with ear linked (A1-A2) reference montage, and a sampling rate of 500 Hz. "Depression, Anxiety and Stress Scale - 21 items" (DASS-21) was used as scoring scheme to generate expert verified ground truth labeling.

2) Dataset 2 (D2): This is a publicly available dataset[1] collected with 19 electrodes referenced with linked ear and placed according to the 10-20 international system [15] [16]. Data was collected from 34 MDD subjects (17 males and 17 females, mean age 40.3 $\pm$ 12.9 years) and 30 healthy controls (21 males and 9 females, mean age 38.3 $\pm$ 15.6 years). MDD subjects were diagnosed by using DSM-IV [17]. The EEG data was acquired in EO and EC states for 5 minutes each. The sampling rate was set to 256 Hz.

B. EEG Data Preprocessing

The raw EEG signal is contaminated with physiological (eye blinks and/or eye movements, muscle or body movements, heartbeats, etc.) and non-physiological (power-line interference, electrode displacement, device error, etc.) artifacts. Removing these artifacts is crucial to increase the identification performance on underlying EEG signals. The preprocessing steps in this work consist of filtering, artifact removal with Independent Component Analysis (ICA) [18], and segmentation.

The EEG data was band-pass (0.5-70 Hz) filtered using a 4th order Chebyshev type II filter. To get rid of the power-line interference, a notch filter at 50 Hz was used. After these filtering operations, the data still contained artifacts such as eye blinks, eye movements, heartbeat, and muscle movements, which were minimized using ICA, where it divides the EEG signals into independent components, and then each component was visually analyzed and removed. The artifact minimized EEG signal was then reconstructed by isolating the undesired component(s).

C. Data segmentation

In the model development process, it is common to divide the EEG signal into shorter segments to speed up the feature extraction step and to increase the amount of training data. In related works, various data lengths were used, which varies between 2, 3, 6, 30, and 75 seconds [19]–[24]. In this work, different segmentation lengths (2, 4, 10, and 20 seconds) were used for ML models. For the DL model, a 1-second segment was selected so that the training data size could be large enough to enable model training. A segment length of $<$1 second might not be informative enough to contain specific characteristics for identification. The total number of data segments for each segmentation length are shown in Table I.

TABLE I
DATA SEGMENTS FOR DIFFERENT SEGMENTATION LENGTHS.

Seg. len. (s)	MDD	Healthy
1	3445	4589
2	1803	2372
4	910	1194
10	364	482
20	194	255

D. Feature extraction

Feature extraction is used for revealing hidden patterns from the finite length EEG signal, say, $\mathbf{x} = \{x(1), x(2), \ldots, x(N)\}^T$, where N is the length of the signal. In this work, linear and non-linear features were extracted based on the most discriminative features proposed by relevant works. Those features were extracted by using 3 [25], 6 [19], 19 [26], or 31 EEG channels. Choice of the channels is based on the literature, and the electrode locations for each dataset are provided in Table II. Since the electrode montages in D1 and D2 were not exactly the same, proximity based closeness was considered to select the channels.

1) Statistical features: Statistical features include mean, variance, skewness, kurtosis, energy, and Hjorth parameters. Hjorth parameters - activity (h_0), mobility (h_1), and complexity (h_2) - are statistical time-domain properties commonly used in feature extraction of biomedical signals [27]. Here, h_0 is the variance and h_1 (Eq.(1)) is the mean frequency of the signal, $\mathbf{x}$.

[1] https://figshare.com/articles/dataset/EEG_Data_New/4244171

TABLE II
EEG CHANNELS SELECTED FROM D1 AND D2

# of ch	Dataset	Channels
3	1	Fp1, Fpz, Fp2
	2	Fp1, Fz, Fp2
6	1	FT7, FT8, T6, T5, TP7, TP8
	2	F7, F8, T6, T5, T3, T4
19	1	Fp1, Fp2, F3, F4, F7, F8, Fpz, TP7, TP8, T5, T6, P3, P4, FT7, FT8, O1, O2, C3, C4
	2	Fp1, Fp2, F3, F4, F7, F8, Fz, T3, T4, T5, T6, P3, P4, Pz, O1, O2, C3, C4, Cz
31	1	Fp1, Fpz, Fp2, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T3, C3, Cz, C4, T4, TP7, CP3, CPz, CP4, TP8, T5, P3, Pz, P4, T6, O1, Oz, O2
	2	—

h_2 (Eq.(1)) is the estimation of signal bandwidth and indicates how similar the signal is to a sine wave.

$$h_1 = \left(\frac{\text{var}\left(\frac{dx(t)}{dt}\right)}{h_0} \right)^{\frac{1}{2}}, \; h_2 = \frac{h_1\left(\frac{dx(t)}{dt}\right)}{h_1(x(t))}. \tag{1}$$

2) Band powers: Using Welch method, the EEG band power is calculated for each distinct frequency bands, i.e., δ (0.5–4 Hz), θ (4–8 Hz), α (8–16 Hz), β (16–32 Hz), and γ (32–70 Hz) [28].

3) Entropy: Entropy measures the uncertainty or randomness in a signal [29]. Sample entropy ($\mathcal{S}_{en}$) defined in Eq.(2) and power spectral entropy ($\mathcal{PS}_{en}$) defined in Eq.(3) are used in this work.

$$\mathcal{S}_{en} = \log\left(\frac{A(s,r)}{A(s+1,r)}\right), \tag{2}$$

where s is the segment length, and r is the threshold for similarity (Chebyshev distance). Each segment is compared to the rest of the segments, and the number of segments within a similarity threshold (r) are summed to obtain A. Higher values indicate that the signal is irregular or random, whereas lower values indicate regularity or repetitiveness. The following values were used in this work: $s = 2$ and $r = 0.2\text{std}(\mathbf{x})$. $\mathcal{PS}_{en}$ is computed as follows:

$$\mathcal{PS}_{en} = -\sum_{f=0}^{f_s/2} P(f)\log_2(P(f)), \tag{3}$$

where $P(f)$ is the normalized power spectral density obtained using Welch method and f_s is the sampling rate.

4) Hurst exponent ($\mathcal{H}_e$): It is a measure of the long-term memory in a time series [30]. $\mathcal{H}_e$ (Eq.(4)) can vary between 0 and 1 and based on its values, the time series can be classified into: (i) Anti-persistent time series (mean-reverting) ($\mathcal{H}_e < 0.5$), (ii) Random walk (impossible prediction) ($\mathcal{H}_e = 0.5$), and (iii) Persistent time series (trending) ($\mathcal{H}_e > 0.5$), where an increase in value will most likely be followed by an increase in the short term memory and vice versa.

$$\mathcal{H}_e = \frac{\log(R/\sigma)}{\log(N/2)}, \tag{4}$$

where R, σ, and N are the range, standard deviation, and length of the time series, respectively.

5) Fractal dimension (FD): It is a measure that quantifies the complexity or self-similarity of the signal. In this work, two different algorithms, namely, Higuchi's and Katz's algorithms were used to estimate the FD.

According to Higuchi's FD (HFD) algorithm [31], if we are given a finite time-series, $\mathbf{x}$ of length N and a parameter $k_{max} \geq 2$, for each $k \in \{1, 2, \ldots, k_{max}\}$ and $m \in \{1, 2, \ldots, k\}$, the length of the curve, $L_m(k)$, is given by:

$$L_m(k) = \frac{N-1}{\lfloor \frac{N-m}{k} \rfloor k^2} \sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} |x(m+ik) - x(m+(i-1)k)|, \tag{5}$$

and $L(k) = \frac{1}{k}\sum_{m=1}^{k} L_m(k)$. In the above equation, $\lfloor \frac{N-m}{k} \rfloor$ is the integer part of the ratio. Finally, the HFD of $\mathbf{x}$ is the slope of a best-fitting line on a plot of $\log\left(\frac{1}{k}\right)$ *vs* $\log L(k)$. In this study, k_{max} was set to 10.

Katz FD (KFD) is another computationally less demanding method of estimating FD [32]. For the time series $\mathbf{x}$, the maximum distance (d) of the data points from the first data point, $x(1)$, is calculated as $d = \max(|x(1) - x(j)|)$, where $j \in \{2, 3, \ldots, N\}$. Then the total length of the time series is calculates as,

$$L = \sum_{i=2}^{N} |x(i) - x(i-1)|. \tag{6}$$

The average distance between two successive points is $a = L/(N-1)$. KFD is then calculated as,

$$\text{KFD} = \frac{\log(L/a)}{\log(d/a)}. \tag{7}$$

6) Detrended fluctuation analysis (DFA): This feature measures the self-affinity and long-term memory of a time series similar to $\mathcal{H}_e$. The advantage of this method is that it can be applied to non-stationary signals, such as EEG signals. To calculate DFA, we first integrate the time series for each value of k ($1 \leq k \leq N$), after subtracting the mean, $\bar{\mathbf{x}}$ of $\mathbf{x}$, to obtain $y(k) = \sum_{i=1}^{k} x(i) - \bar{\mathbf{x}}$. The integrated time series is then segmented into smaller sections of length m. In each of these segments, a least squares line is fit to the data (representing the trend in that segment), denoted as $y_m(k)$. The next step is to detrend the integrated time series by subtracting the local trend, $y_m(k)$. The fluctuation is then calculated as,

$$F(m) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y(k) - y_m(k))^2}. \tag{8}$$

Finally, DFA is computed as the slope of a straight line fit to the double logarithmic ("log-log") graph of $F(m)$ against m, frequently referred to as a fluctuation plot [33].

E. Feature selection

In the model building pipeline, feature selection is a crucial step in identifying the subset of relevant input features. The feature set can contain redundant and noisy data, hindering the interpretability of the model. By performing feature selection,

it is possible to reduce over-fitting, improve accuracy, and speed up the model training time. In this work, the analysis of variance (ANOVA), genetic algorithm (GA), and minimum redundancy maximum relevance (mRMR) methods are used for feature selection.

1) ANOVA: It is a statistical test for checking if the means of two or more samples differ. ANOVA calculates the F-value for each feature, which indicates if the means for different classes are statistically and significantly different or not. This could be used to select the best features with significantly different mean values between the two classes. For example, if the $\mathcal{H}_e$ is considered and it has a high F-value close to 1, then the mean of $\mathcal{H}_e$ is different for depressed and healthy subjects. On the other hand, if F-value is low (≈ 0), then the mean of $\mathcal{H}_e$ is similar for both classes suggesting it is not a good feature.

2) GA: It is a method for solving both constrained and unconstrained optimization problems. For feature selection, first, random subsets of features are selected to create populations. Each population is evaluated with the predictive model as a task at hand, and evolved through generations by crossover (combining successful features) and mutation (introducing randomly selected new features).

3) mRMR: In this method [34] of feature selection, at every iteration (i), the algorithm selects the features that have a high correlation with the target variable and a low correlation with features that have already been selected in the previous iterations. Thus, the score for each feature (z) is:

$$\text{score}_i(z) = \frac{\text{relevance}(z|\ \text{target})}{\text{redundancy}(z|\ \text{features selected})}. \tag{9}$$

The algorithm adds feature with the highest score to the selected feature set at each iteration. The relevance is calculated as F-statistic between the feature and the target, while redundancy is the average of Pearson correlations of feature z and features selected at previous iterations.

F. Classifiers

The classification problem concerns a supervised learning task, as the algorithms are trained on labeled data and there are two classes, either healthy or depressed. In the present study, for the classification task, k-nearest neighbors (k-NN) [35], support vector machine (SVM) [36], random forest (RF) [37], extreme gradient boosting (XGBoost) [38] and an ensemble of them were applied.

1) k-Nearest Neighbors: It is a simple algorithm for performing supervised classification. The algorithm operates on the principle that samples belonging to the same class are clustered in the feature space. To classify a new data point, the k-NN algorithm calculates the distances between this new point and its nearest neighbors and then selects the k closest neighbors. The final step involves assigning the new data point to a class through a majority vote of the selected k neighbors.

2) Support Vector Machine: SVM operates by finding the optimal hyperplane in a multi-dimensional feature space that separates the data points into two distinct classes. The hyperplane is optimized to maximize the margin, or distance, from the nearest data points from either class. When the data is not linearly separable, a kernel can be used to transform the data into a linearly separable space. In this study, a polynomial kernel was used as suggested by grid search.

3) Random Forest: It is a ML algorithm that uses a collection of decision trees to make predictions. The algorithm operates by having each individual decision tree make a prediction, with the final prediction being determined by the class that receives the most votes. Diversity of the decision trees is ensured through the random subset of features used in their training, leading to a more robust overall model. This combination of individual diverse models has been shown to improve prediction accuracy compared to using any of the individual model alone.

4) Extreme Gradient Boosting: XGBoost is a decision tree-based algorithm that utilizes the boosting technique. Unlike RF, which uses a bagging technique where decision trees work in parallel, XGBoost combines decision trees sequentially. The subsequent trees aim to correct the mistakes made by previous trees by focusing on samples that the previous trees mis-classified. This sequential combination of decision trees results in a highly accurate and precise model.

5) Ensemble technique: The combination of multiple individual model predictions to make a single prediction is achieved by using the ensemble model. There are various techniques for combining multiple models, such as averaging predictions, weighted averaging, or majority voting. In this work, the outputs from $k-$NN, SVM, RF, and XGBoost models are combined using the Dempster-Shafer combination rule [39].

III. Proposed Convolutional Neural Network Architecture and Training

In this section, a theoretical description of Convolutional Neural Network (CNN) is presented. Topology of the proposed deep CNN is also outlined with a description of the training procedure.

A. Convolutional Neural Network

CNN is a type of DL algorithm with superior performance compared to other DL models on images and audio data. They typically consist of convolutional, pooling, fully connected, dropout and batch normalization layers.

1) Convolutional Layer: Convolutional layer is the core building block of CNN. It involves an input and kernel or feature detector. The kernel is aligned to an area of the input, and the dot product is performed. The kernel then repeats the procedure with shifting by stride. One convolutional layer can contain many kernels and the number of kernels affects the depth of an output. After each convolutional layer, activation functions, such as ReLU or hyperbolic tangent, is applied to introduce non-linearity.

2) Pooling Layer: The pooling layer reduces the dimensionality of the feature map, leading to quicker computation and fewer parameters to be learned. Like the convolutional layer, it employs kernels, however, here the kernel selects either the maximum (max-pooling) or mean (average pooling) values within the receptive field.

3) Fully Connected Layer: Fully connected layer is typically the final layer in the CNN. As the name suggests, all nodes from the previous layer are connected to all nodes in the next layer. The fully connected layer takes the input features generated in previous layers and performs classification based on them. It is also usually followed by a non-linear activation function, except for the last layer.

4) Dropout and Batch Normalization: Dropout randomly drops (sets to zero) the output of neurons with a certain probability. If neurons are randomly dropped, other neurons have to make predictions, resulting in a network consisting of independent neurons that do not rely on a specific previous or neighbouring neuron. Batch normalization is another technique that makes the training more stable and faster by normalizing the input to each layer.

B. Proposed Architecture

Figure 1 illustrates the proposed architecture of the CNN model. EEG signals were processed the same way as it was for training ML models, i.e., filtered and ICA performed. The network takes as input a one second length of EEG signal from all 31 channels. The data is fed to 3 blocks consisting of convolutional, max-pooling, batch normalization, and dropout layers, followed by a fully-connected layer and output layer. The kernel size is chosen to be 11×7, as it was observed that a larger kernel size gives better performance. The addition of batch normalization and dropout with a 50% rate help to prevent overfitting. The activation function is chosen as the Leaky ReLU.

C. Training

The network was built and trained using Keras library. The model was trained for 200 epochs with a batch size of 4. The learning rate was set to 0.0005 with a cosine decay to reduce the learning rate as the training progressed. As an optimizer, a computationally efficient and low memory-demanding Adam algorithm was used [40]. The distribution between train, test, and validation sets is chosen as 70:15:15.

IV. Results and Discussion

In this section, results obtained with varying channel numbers, segmentation lengths, and feature selection algorithms are presented. All presented results for ML models are for 10-fold cross-validation.

A. Effect of number of channels

To identify an optimal number of channels for the diagnosis of depression, the performance of five different classifiers was tested on various channel numbers. The segmentation length was fixed to be 10 seconds and all other parameters varied. The summary of various parameters is given in Table III. Table IV provides the results (accuracy) obtained for a 10-fold cross-validation. Results are listed for each classifier's best-performing feature selector and the percentage of selected features.

From the results, it is observed that features extracted from 31 channels give the best results when only a quarter of the features are selected with the GA. The highest accuracy of $99.62 \pm 0.58\%$ is obtained with the ensemble model. However, for 31 channel EEG classification, the accuracies of other classifiers are also high ($>98\%$). It is observed that the 3 channel case has the highest accuracy ($92.33 \pm 2.21\%$ with the XGBoost classifier and 90% features selected with mRMR). In general, as the number of channels increases, there is also an increase in the model performance. Additionally, the ensemble model performs better for all cases except 3-channel case where XGBoost performs better.

B. Effect of segmentation length

The effect of segmentation length on model performance was tested. Selected segmentation lengths are 2, 4, 10, and 20 seconds. The results are provided for 19-channel case. ANOVA was used for feature selection with 90% of features selected. Performance metrics of the five classifiers for various segmentation lengths are shown in Table V. It can be observed that the model accuracies are approximately similar for all segmentation lengths and all are generally high, $>95\%$. However, for the 10-second case, all models have an accuracy of 98% or higher and thus variance between model performances is lower compared to other cases. For instance, in the 2-second case, the model accuracies vary between 98.69% to 95.12%. Also, it should be noted that most of the models (k-NN, RF and Ensemble) reached their highest accuracies when 10-second segments are used, whereas SVM and XGBoost with 20 and 2-second segments, respectively.

C. Effect of feature selection method

For comparison of different feature selection methods, 19 channels were used with a segmentation length of 10 seconds and 90% of features selected.

A comparison plot of all three feature selectors for various models is shown in Fig. 2. It can be seen that the GA performs better with 25% of features for all classifiers. However, it gets outperformed by ANOVA and mRMR when more features are selected. ANOVA and mRMR have around the same performance. Both methods are model agnostic, meaning they can be used with any type of predictive model and do not require the model to choose features. Also, both methods utilize F-values as a measure of the significance of differences between groups. Therefore, it is reasonable that they have the same performance. Furthermore, it is noteworthy that GA demonstrates superior performance in comparison to other feature selection techniques when applied to ensemble model.

TABLE III
OPTIMIZED HYPERPARAMETERS AND REFERENCE TO THE SECTION

	Options	**Optimized section**
Channels	3, 6 , 9 , 31	IV - A
Feature percentage	25, 50, 75, 90 , 100	IV - A
Segmentation length (s)	2, 4, 10 , 20	IV - B
Feature selector	ANOVA , GA, mRMR	IV - C

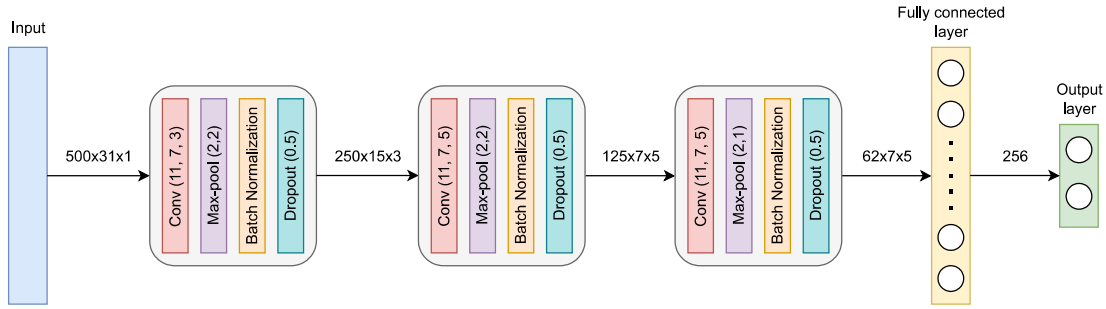


Fig. 1. Employed CNN architecture for subject classification.

TABLE IV
MODEL INFERENCE RESULTS ON TEST DATA FOR VARIOUS CHANNEL NUMBERS.

# of ch	Model	Feature selector (% of features)	# of features	Accuracy	Precision	Recall	F1-score
31	k-NN	All features	558 of 558	99.12 ± 0.80	100.00 ± 0.0	97.76 ± 2.06	98.51 ± 1.59
	SVM	ANOVA (90%)	502 of 558	99.12 ± 0.98	97.92 ± 3.69	97.76 ± 3.24	97.91 ± 1.27
	RF	mRMR (75%)	418 of 558	98.49 ± 1.09	99.04 ± 2.02	95.86 ± 2.85	96.70 ± 2.70
	XGBoost	mRMR (75%)	418 of 558	98.74 ± 0.98	98.12 ± 2.81	95.54 ± 4.06	97.21 ± 1.82
	Ensemble	GA (25%)	90 of 558	99.62 ± 0.58	99.69 ± 0.94	99.35 ± 1.29	99.51 ± 1.05
19	k-NN	ANOVA (75%)	256 of 342	98.56 ± 1.50	99.18 ± 1.74	97.14 ± 2.56	97.68 ± 1.18
	SVM	ANOVA (90%)	307 of 342	97.84 ± 1.85	98.29 ± 1.85	96.30 ± 2.58	97.55 ± 1.84
	RF	All features	342 of 342	98.09 ± 1.43	99.15 ± 1.30	95.45 ± 2.58	97.25 ± 1.92
	XGBoost	mRMR (90%)	307 of 342	98.44 ± 1.43	98.66 ± 3.12	97.17 ± 2.84	97.68 ± 2.27
	Ensemble	GA (90%)	106 of 342	99.28 ± 1.09	99.20 ± 1.70	99.14 ± 1.31	98.86 ± 0.86
6	k-NN	GA (50%)	25 of 108	97.36 ± 1.19	98.86 ± 1.87	96.02 ± 1.89	96.69 ± 1.84
	SVM	GA (90%)	37 of 108	93.64 ± 2.35	95.60 ± 3.30	90.05 ± 5.00	92.24 ± 3.07
	RF	All features	108 of 108	96.53 ± 2.23	96.85 ± 2.81	94.02 ± 2.98	95.64 ± 1.68
	XGBoost	ANOVA (75%)	81 of 108	97.01 ± 1.62	96.83 ± 2.73	96.01 ± 3.18	95.98 ± 2.12
	Ensemble	ANOVA (90%)	97 of 108	98.44 ± 1.07	98.54 ± 1.99	96.03 ± 3.15	97.10 ± 2.11
3	k-NN	ANOVA (90%)	48 of 54	85.48 ± 3.55	91.11 ± 5.83	71.87 ± 6.54	81.23 ± 2.34
	SVM	mRMR (75%)	40 of 54	78.53 ± 3.79	75.10 ± 7.51	69.60 ± 3.59	72.39 ± 8.47
	RF	ANOVA (75%)	40 of 54	90.16 ± 3.84	88.61 ± 5.20	84.67 ± 5.26	87.31 ± 3.97
	XGBoost	mRMR (90%)	48 of 54	92.33 ± 2.21	90.40 ± 5.89	89.46 ± 7.25	90.94 ± 4.03
	Ensemble	ANOVA (90%)	48 of 54	91.25 ± 3.50	90.53 ± 5.10	90.05 ± 2.95	89.65 ± 3.68

Segmentation length is fixed to be 10 sec. The metrics are with mean and standard deviation for 10-fold cross-validation runs.

D. DL method

The model's performance on test data (D1) is given in Fig. 3. According to the confusion matrix, the model has an accuracy of 98.74% with almost the same performance for both classes.

TABLE V
RESULTS FOR VARIOUS SEGMENTATION LENGTHS

Seq. len (s)	Model	Accuracy	Precision	Recall	F1-score
2	k-NN	96.69 ± 0.70	98.91 ± 0.92	93.58 ± 2.85	96.13 ± 0.69
	SVM	95.12 ± 1.19	94.85 ± 1.42	93.52 ± 2.23	94.13 ± 1.31
	RF	97.45 ± 0.58	97.96 ± 0.87	96.39 ± 1.45	96.70 ± 1.19
	XGBoost	98.69 ± 0.52	98.80 ± 0.78	97.48 ± 1.12	98.36 ± 0.70
	Ensemble	98.52 ± 0.53	98.56 ± 0.99	97.88 ± 0.68	98.39 ± 0.66
4	k-NN	97.83 ± 0.95	98.99 ± 1.54	95.57 ± 2.41	97.63 ± 0.99
	SVM	97.06 ± 1.17	96.71 ± 1.26	96.14 ± 1.54	96.63 ± 1.21
	RF	97.88 ± 1.10	98.29 ± 1.52	96.70 ± 2.12	97.06 ± 1.18
	XGBoost	98.55 ± 0.75	97.97 ± 1.85	97.84 ± 1.29	97.71 ± 0.74
	Ensemble	98.75 ± 0.81	98.43 ± 1.52	98.41 ± 1.27	98.58 ± 0.93
10	k-NN	98.44 ± 0.93	98.90 ± 1.35	96.58 ± 2.80	98.11 ± 1.89
	SVM	97.84 ± 1.85	98.29 ± 1.85	96.30 ± 2.58	97.55 ± 1.84
	RF	98.08 ± 1.10	98.34 ± 2.19	97.71 ± 2.80	98.00 ± 0.93
	XGBoost	98.32 ± 1.71	98.03 ± 2.20	96.60 ± 3.04	97.83 ± 1.63
	Ensemble	99.04 ± 1.04	98.59 ± 1.88	98.86 ± 1.40	98.72 ± 0.99
20	k-NN	96.39 ± 3.38	98.91 ± 2.19	92.02 ± 5.41	96.42 ± 3.10
	SVM	98.43 ± 2.25	99.44 ± 1.67	95.23 ± 4.97	97.52 ± 1.99
	RF	96.38 ± 3.55	97.47 ± 3.31	93.10 ± 6.29	96.42 ± 2.88
	XGBoost	97.06 ± 2.29	95.86 ± 5.05	94.65 ± 4.78	94.12 ± 5.77
	Ensemble	97.74 ± 2.49	97.50 ± 4.03	97.87 ± 2.62	98.14 ± 1.70

Fixed channels numbers (19), percentage of features selected (90%), feature selection methods (ANOVA). The metrics are shown with mean and standard deviation for 10-fold cross-validation runs

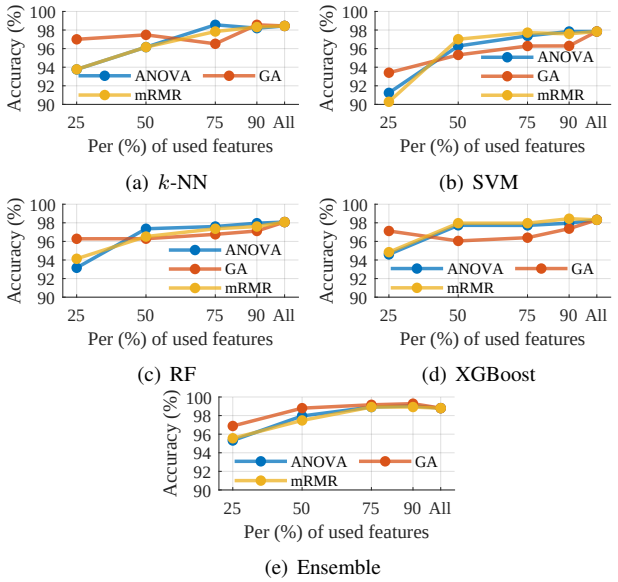


Fig. 2. The average accuracy of (a) k-NN, (b) SVM, (c) RF, (d) XGBoost and (e) Ensemble model for various percentages of features chosen (25%, 50%, 75%, 90%, and all features) and feature selector methods (ANOVA, GA, mRMR)

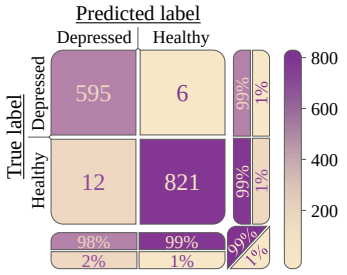
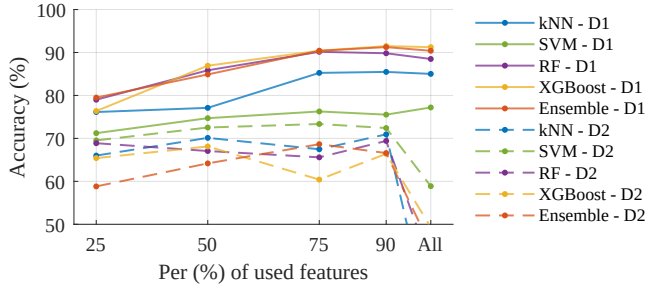


Fig. 3. Confusion matrix for test data (D1).

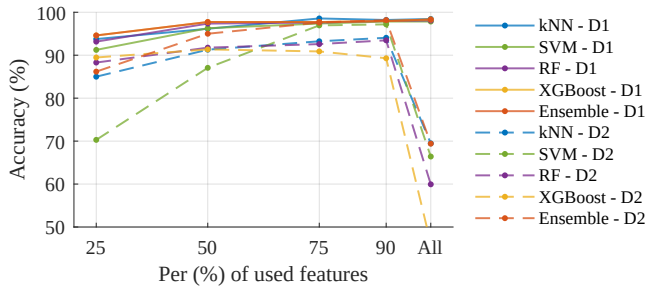
E. Generalization test

The trained models are expected to be generalizable to perform well on unseen data. Both ML and DL models have good accuracy on unseen test data achieving an accuracy above 98%. However, in this work, models trained on D1 will be tested on D2 to see if they can generalize on other datasets. It is a complicated task as there are differences between datasets regarding EEG devices, channels, sampling frequency, background noise, montage, and preprocessing steps applied during data acquisition.

Figure 4(a) shows the performance of ML models on D1 and D2 for various percentages of features used for 3 channels case. Referring to the figure, XGBoost and k-NN have the highest difference in performance on both datasets, with a difference of approximately 10 to 30%. On the other hand, SVM has the best ability to generalize with a variation of only around 3% between datasets for cases except when all features are used. This figure (Fig.4(b)) also shows model performances for 19-channel cases. It is observed that the models can generalize even better when more channels are used. Also, the ensemble model has a better performance compared to others on D2



(a) Test for generalization for 3 channels case



(b) Test for generalization for 19 channels case

Fig. 4. ML model generalization performance on D1 and D2 for various percentages of features extracted from 3 (a) and 19 (b) channels with 10-second segmentation length and ANOVA feature selection technique.

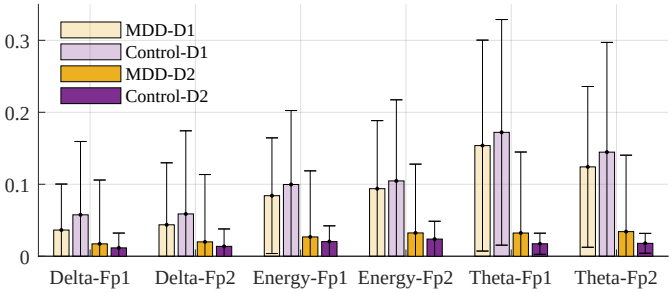


Fig. 5. Six features negatively affecting the generalization ability of the models.

when 19 channels are used.

Interestingly, when all of the features are used, the ability to generalize for all models drops dramatically, as low as 26% for k-NN. To find an explanation for this trend, a closer analysis was performed on those features that are not included in the 90% of features extracted in the 3 channels (Fp1, Fp2 and Fpz) case. When all features are used, there are 54 features in total and when 90% of features are selected, the model is trained on 48 features. This indicates that 6 new features cause a significant reduction in model performance on D2. The mean and variance of those 6 features for D1 and D2 are shown in Fig. 5. As illustrated for those 6 features, an opposite relation exists between for MDD and control group subjects between D1 and D2.

The generalization ability of the CNN model was also tested on D2, which has only 19 channels, whereas CNN presented before was trained for 31 channels. Thus, the model was retrained on D1 with 19 channels for testing on D2 with the same architecture and hyper-parameters. The final model achieved an accuracy of 93.7% on test data from D1, but failed to generalize on D2 with an accuracy of only 43.7%. However, when trained on D2, the same model achieves an accuracy of 97.3% on a test data from D2.

F. Discussion

One of the main objectives of this work is to study the effect of number of channels, segmentation length, and feature selection techniques. From the experiments with varying number of channels (Table IV), it is evident that an increase in the number of channels results in higher accuracy, with the highest accuracy of $99.56 \pm 0.58\%$ achieved for 31 channel case with the ensemble model. However, having more channels increases computational time to extract features from each of those channels. Therefore, the number of channels should be selected for the specific application depending on requirements for inference time and accuracy.

Another hyper-parameter that can impact the model is the segmentation length. Experiments performed with segmentation lengths of 2, 6, 10, and 20 seconds suggest that there is not much of a difference in model performance, with all of them having high accuracies with slight variation. However, the 10-second case has better stability across different classifiers.

This study has also compared ANOVA, GA, and mRMR feature selection methods. For SVM, RF, and XGBoost, reducing the feature set by half with ANOVA or mRMR gives

comparable results to the case when all features are used. Additionally, the reduction of the feature set further results in a sharper negative impact on accuracy. Even though GA performs worse when the feature set is 50% or higher, it performs better when only 25% of features are used. Thus, this experiment suggests that GA should be the choice if a significant reduction in the feature set is required. Otherwise, mRMR or ANOVA are better choices with significantly less computational requirements. In addition, although feature selection does not result in a significant increase in performance, it improves the ability to generalize, as illustrated in Fig. 4.

In terms of generalisation, all trained classifiers can generalize well on unseen data from the same dataset with a similar data acquisition procedure. To assess generalization further, the models trained with D1 were tested on D2. All ML models are able to perform better than random guess staying above 60% accuracy when 90% or less features are selected. ML models can generalize much better when 19 channels are used compared to when fewer channels are used. Also, results suggest that when building generalizable models, it is essential to perform feature selection. Otherwise, the performance could drop below random guess. It is also important to compare the feature distributions (mean, variance, etc.) between datasets per class to ensure the consistency.

Even though ML models can generalize well, CNN has poor performance on generalization. This could be attributed to the fact that the CNN has learned features that are specific to D1 but not present in D2 or that CNN requires more extensive, more diverse training data to generalize well. Further analysis is suggested to find an explanation for this and further examination of the interpretability of the models is necessary.

When comparing ML and DL classification methods, we must focus on performance, simplicity, speed, generalization ability, and explainability. ML models, particularly the ensemble model attains slightly higher accuracy than DL method, 99.62% versus 98.74%. DL methods have the advantage of scaling and improving as the data quality and quantity increases. Another advantage of DL methods is that they are simpler than ML methods because they do not require manual feature extraction and selection, they learn important characteristics directly from data. Even in some works, the CNN model takes the raw EEG data without any preprocessing or data cleaning [41]. This makes DL model development and inference time faster than ML methods. Tests were performed on the inference time of ML and DL methods. The inference time of ML models on 19 channel data, including extracting features for each channel, is 27.45s, whereas the inference time for the CNN model is 1.14s. Thus, there is a huge difference in inference time between the methods. As discussed before, concluding that ML methods are more stable on varying dataset. Another point is explainability, which has critical importance in the medical field. Classical ML methods definitely have an advantage of explainability compared to DL models. But, with various techniques such as Grad-CAM, SHAM, or LIME behaviour of models can be explained [42]–[44].

V. Conclusion and Future Work

The intersection of AI and EEG holds significant potential to revolutionize the diagnosis and treatment of depression. One of the significant challenges in diagnosing depression is the subjective nature of self-reporting symptoms. AI and EEG integration can potentially overcome this challenge by providing objective measurements of brain activity, which can serve as a complementary diagnostic tool.

The main aims of this research were to propose an accurate way of diagnosing depression, study the effect of various hyper-parameters systematically and compare ML and DL based methods. An ensemble model that combines the predictions of k-NN, SVM, RF and XGBoost with the Dempster-Shafer rule produced the highest accuracy of $99.62 \pm 0.58\%$ for 31 channels and 10-second segmentation lengths. CNN model has also provided a high accuracy of 98.74% for 1-second EEG signals from 31 channels. The generalization test results suggest that ML models can generalize well without training on D2 when a larger number of features are used and combined with feature selection methods. However, CNN model fails to generalize on another dataset which could be attributed to the limited dataset size.

The study on the effect of hyper-parameters indicates that number of channels, feature selection method, and the number of features selected significantly impact model performance. In contrast, the segmentation length does not seem to be as important. As features are extracted from larger numbers of channels, the model performance improves, but there is also a trade-off between feature extraction time and accuracy. The GA tends to yield better results with a smaller subset of features (25%) and also works well in combination with the ensemble model. By comparing ML and DL methods, it can be outlined that DL methods are faster and simpler as it has inherent feature extraction and selection capabilities.

Future research should aim to verify our findings with larger and more diverse data to validate these results. There is a need for public datasets with a variable age range, demographics, noise levels, etc. Also, there is a demand for an automatic artifact removal method that can reliably remove as many of them as possible (muscle, eye movement, blinks, heartbeat, etc.).

References

[1] "Depression," https://www.who.int/news-room/fact-sheets/detail/depression, accessed: 2022-10-11.
[2] "World Mental Health Day: Depression to be the biggest cause of ill health by 2030, says WHO," https://www.indiatoday.in/education-today/gk-&-current-affairs/story/world-mental-health-day-depression-suicide-prevention-who-1607866-2019-10-10, accessed: 2022-10-11.
[3] "International Classification of Diseases,Tenth Revision (ICD-10)," https://www.cdc.gov/nchs/icd/icd10.htm, accessed: 2022-10-11.
[4] D. American Psychiatric Association, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.
[5] S. D. Østergaard, S. Jensen, and P. Bech, "The heterogeneity of the depressive syndrome: when numbers get serious." *Acta Psychiatrica Scandinavica*, 2011.
[6] W. Liu, K. Jia, Z. Wang, and Z. Ma, "A depression prediction algorithm based on spatiotemporal feature of EEG signal," *Brain Sciences*, vol. 12, no. 5, p. 630, 2022.

[7] A. Dev, N. Roy, M. K. Islam, C. Biswas, H. U. Ahmed, M. A. Amin, F. Sarker, R. Vaidyanathan, and K. A. Mamun, "Exploration of EEG-based depression biomarkers identification techniques and their applications: A systematic review," *IEEE Access*, 2022.

[8] R. Kerestes, C. G. Davey, K. Stephanou, S. Whittle, and B. J. Harrison, "Functional brain imaging studies of youth depression: a systematic review," *NeuroImage: Clinical*, vol. 4, pp. 209–231, 2014.

[9] J. Persson, A. Wall, J. Weis, M. Gingnell, G. Antoni, M. Lubberink, and R. Bodén, "Inhibitory and excitatory neurotransmitter systems in depressed and healthy: A positron emission tomography and magnetic resonance spectroscopy study," *Psychiatry Research: Neuroimaging*, vol. 315, p. 111327, 2021.

[10] S. Mahato and S. Paul, "Electroencephalogram (EEG) signal analysis for diagnosis of major depressive disorder (MDD): a review," *Nanoelectronics, Circuits and Communication Systems*, pp. 323–335, 2019.

[11] S. Olbrich and M. Arns, "EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response," *International Review of Psychiatry*, vol. 25, no. 5, pp. 604–618, 2013.

[12] O. Faust, P. C. A. Ang, S. D. Puthankattil, and P. K. Joseph, "Depression diagnosis support system based on EEG signal entropies," *Journal of mechanics in medicine and biology*, vol. 14, no. 03, p. 1450035, 2014.

[13] G. M. Bairy, U. C. Niranjan, and S. D. Puthankattil, "Automated classification of depression EEG signals using wavelet entropies and energies," *Journal of Mechanics in Medicine and Biology*, vol. 16, no. 03, p. 1650035, 2016.

[14] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Computer methods and programs in biomedicine*, vol. 161, pp. 103–113, 2018.

[15] W. Mumtaz, L. Xia, M. A. Mohd Yasin, S. S. Azhar Ali, and A. S. Malik, "A wavelet-based technique to predict treatment outcome for major depressive disorder," *PloS one*, vol. 12, no. 2, p. e0171409, 2017.

[16] G. H. Klem, "The ten-twenty electrode system of the international federation: the international federation of clinical neurophysiology," *Electroencephalogr. Clin. Neurophysiol. Suppl.*, vol. 52, pp. 3–6, 1999.

[17] C. C. Bell, "DSM-IV: diagnostic and statistical manual of mental disorders," *Jama*, vol. 272, no. 10, pp. 828–829, 1994.

[18] A. Tharwat, "Independent component analysis: An introduction," *Applied Computing and Informatics*, vol. 17, no. 2, pp. 222–249, 2021.

[19] S. Mahato, N. Goyal, D. Ram, and S. Paul, "Detection of depression and scaling of severity using six channel EEG data," *Journal of medical systems*, vol. 44, no. 7, pp. 1–12, 2020.

[20] A. Seal, R. Bajpai, M. Karnati, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, and O. Krejcar, "Benchmarks for machine learning in depression discrimination using electroencephalography signals," *Applied Intelligence*, pp. 1–18, 2022.

[21] Y. Mohammadi, M. Hajian, and M. H. Moradi, "Discrimination of depression levels using machine learning methods on EEG signals," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2019, pp. 1765–1769.

[22] J. Zhu, Y. Wang, R. La, J. Zhan, J. Niu, S. Zeng, and X. Hu, "Multimodal mild depression recognition based on EEG-EM synchronization acquisition network," *IEEE Access*, vol. 7, pp. 28 196–28 210, 2019.

[23] Y. Li, Y. Shen, X. Fan, X. Huang, H. Yu, G. Zhao, and W. Ma, "A novel EEG-based major depressive disorder detection framework with two-stage feature selection," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–13, 2022.

[24] S. Saleque, R. I. Kamal, R. T. Khan, A. Chakrabarty, M. Z. Parvez *et al.*, "Detection of major depressive disorder using signal processing and machine learning approaches," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2020, pp. 1032–1037.

[25] J. Shen, S. Zhao, Y. Yao, Y. Wang, and L. Feng, "A novel depression detection method based on pervasive EEG and EEG splitting criterion," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1879–1886.

[26] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Computer methods and programs in biomedicine*, vol. 109, no. 3, pp. 339–345, 2013.

[27] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and clinical neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[28] A. Alkan and M. K. Kiymik, "Comparison of AR and Welch methods in epileptic seizure detection," *Journal of Medical Systems*, vol. 30, pp. 413–419, 2006.

[29] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[30] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American society of civil engineers*, vol. 116, no. 1, pp. 770–799, 1951.

[31] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[32] M. J. Katz, "Fractals and the analysis of waveforms," *Computers in biology and medicine*, vol. 18, no. 3, pp. 145–156, 1988.

[33] R. Bryce and K. Sprague, "Revisiting detrended fluctuation analysis," *Scientific reports*, vol. 2, no. 1, p. 315, 2012.

[34] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[35] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1255–1260.

[36] Y. Zhang, "Support vector machine classification algorithm and its application," in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3*. Springer, 2012, pp. 179–186.

[37] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[38] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[39] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] H. Kwon, S. Kang, W. Park, J. Park, and Y. Lee, "Deep learning based pre-screening method for depression with imagery frontal EEG channels," in *2019 International conference on information and communication technology convergence (ICTC)*. IEEE, 2019, pp. 378–380.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[43] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should I trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.