Effective EEG Feature Selection for Interpretable MDD (Major Depressive Disorder) Classification

Vojtech Mrazek
mrazek@fit.vutbr.cz
Brno University of Technology
Faculty of Information Technology
Brno, Czechia

Soyiba Jawed
jawed@fit.vutbr.cz
Brno University of Technology
Faculty of Information Technology
Brno, Czechia

Muhammad Arif*
muhammad.arif@aau.at
Institute of Networked and Embedded Systems
University of Klagenfurt
Klagenfurt, Austria

Aamir Saeed Malik
malik@fit.vutbr.cz
Brno University of Technology
Faculty of Information Technology
Brno, Czechia

ABSTRACT

In this paper, we propose an interpretable electroencephalogram (EEG)-based solution for the diagnostics of major depressive disorder (MDD). The acquisition of EEG experimental data involved 32 MDD patients and 29 healthy controls. A feature matrix is constructed involving frequency decomposition of EEG data based on power spectrum density (PSD) using the Welch method. Those PSD features were selected, which were statistically significant. To improve interpretability, the best features are first selected from feature space via the non-dominated sorting genetic (NSGA-II) evolutionary algorithm. The best features are utilized for support vector machine (SVM), and k-nearest neighbors (k-NN) classifiers, and the results are then correlated with features to improve the interpretability. The results show that the features (gamma bands) extracted from the left temporal brain regions can distinguish MDD patients from control significantly. The proposed best solution by NSGA-II gives an average sensitivity of 93.3%, specificity of 93.4% and accuracy of 93.5%. The complete framework is published as open-source at https://github.com/ehw-fit/eeg-mdd.

CCS CONCEPTS

• **Applied computing** → *Health informatics*; **Bioinformatics**; • **Computing methodologies** → **Machine learning**; **Bio-inspired approaches**.

KEYWORDS

electroencephalogram (EEG), feature extraction, major depressive disorder

*Also with Brno University of Technology, Faculty of Information Technology, Czechia.

GECCO '23, July 15–19, 2023, Lisbon, Portugal

ACM ISBN 979-8-4007-0119-1/23/07.
https://doi.org/10.1145/3583131.3590398

ACM Reference Format:
Vojtech Mrazek, Soyiba Jawed, Muhammad Arif, and Aamir Saeed Malik. 2023. Effective EEG Feature Selection for Interpretable MDD (Major Depressive Disorder) Classification. In *Genetic and Evolutionary Computation Conference (GECCO '23), July 15–19, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3583131.3590398

1 INTRODUCTION

According to the World Health Organization (WHO), "a state of complete physical, mental, emotional and social well-being of a person and not merely the absence of disease and infirmity corresponds to good health". However, in most parts of the world, physical health is prioritized over mental health leading to an alarming rise in various mental health issues all over the world. For mental health well-being, several mental disorders have been identified such as anxiety, depression, ADHD, schizophrenia etc. Among all of the existing mental health issues, depression is the most prevalent and is projected as a serious threat to mankind. According to the WHO, depression affects more than 264 million people worldwide and is the primary cause of 800 000 suicide deaths each year [31]. Depression can lead to several issues such as sadness, loss of attention or pleasure, demotivation, sleep disorders, exhaustion, and poor attention or focus. In European Union (EU), it is reported that depression affected 7.2% of the EU population in 2018 [27]. In EU the estimated annual direct cost of depression was €620 billion (more than 4% of GDP) in 2018 with costs of €260 billion in unemployment, €190 billion in the health sector (treatment of depression), and €170 billion for the social welfare systems (disability benefits) [20]. Recently the COVID-19 pandemic has further increased the number of depressed individuals [7]. As a result, depression has become a major disease affecting human mental health, and its consequences cannot be ignored. To address this issue, the first step required is the correct diagnosis of depression, followed by treatment and the corresponding efficacy.

For the diagnosis of major depression disorder (MDD) patients, the traditional method used by physicians and psychiatrists is the clinical questionnaire-based assessment, which is primarily determined by patients' responses and behavioral activities [17]. Thus, it is highly susceptible to human subjectivity, which impairs the objectivity of the diagnosis process [17]. Consequently, numerous studies

have been conducted to advance the traditional models' competency as well as develop better replaceable strategies for diagnosing depression. Some of them include computing heart rate variability (HRV), using functional MRI (fMRI), and evaluating through visual facial expression (VFE) [17]. However, these strategies have several disadvantages which limit their practicability in a clinical environment; for example, the outcomes can be significantly affected by body movements in HRV, fMRI is costly and usually not available in smaller clinical settings, and VFE needs long-term and careful monitoring. A good alternative is to use electroencephalography (EEG). EEG records various electrical signals that are generated for communication between brain cells that belong to various parts of the brain. This allows us to learn about different brain behaviors. In addition, EEGs are widely available, mobile, cost-effective, less complex, and patient-friendly. Thus, EEG becomes a powerful tool to analyze and diagnose many mental diseases and disorders such as MDD, schizophrenia, Parkinson's disease, Alzheimer's disease, epilepsy, sleep disorders, and dementia [16]. Furthermore, studies have shown that underline brain neural activities are affected by depression, thus making EEG suitable to study depression [17].

Despite the various advantages, EEG-based classification of MDD patients is still a challenging task and no commercial framework is available for clinicians to diagnose MDD patients using this modality. Therefore, this issue is a hot topic in the research community and researchers are trying to improve the EEG-based diagnosis processes based on accuracy, sensitivity, specificity, and interpretability. The classification of Major Depressive Disorder (MDD) patients from Electroencephalogram (EEG) data traditionally involves two steps. First, a range of feature extraction techniques are used to extract meaningful features from EEG data. These techniques typically involve extracting quantitative and qualitative descriptors of EEG signals. Second, a suitable classification technique is employed on the extracted features to classify MDD patients. Ultimately, the goal of these two steps is to accurately classify MDD patients from their EEG data [16]. Furthermore, based on the extracted features, several classification techniques have been proposed in the literature such as logistic regression (LR), artificial neural networks (ANN), support vector machines (SVM), and convolutional neural networks (CNN) [10]. Because of superior classification performance, Deep learning (DL) techniques have also been proposed for MDD classification [28]. The advantage of DL techniques is that both feature extraction and classification are done by DL simultaneously.

However, the existing traditional ML-based classification solutions use hand-crafted features [6, 18] and thus offer a solution that is difficult to scale, while other solutions using DL provide no interpretability at all [28]. In order to improve the interpretability of the solution, one of the alternatives is the use of genetic algorithms (GAs). GAs have recently attracted many researchers in solving various optimization problems. GAs searches are based on one or more objective functions and have stochastic search abilities in complex environments. This enables them to select the optimal solution among large solutions set for an optimization problem.

The average amount of data generated across many different activities has grown significantly in size as a result of the development of various technologies as well as due to the data processing and storage capabilities of computing devices in the modern era. Thus, to produce better results in various learning tasks, selecting a suitable number and the optimal features from such large datasets is becoming a challenging area of research. To address these issues, GAs have been successfully implemented in the literature for the heuristic selection of the close-to-optimal features by removing unnecessary and redundant features from a high-dimensional feature space [21]. For the feature extraction, some of the popular implemented GAs are ant colony optimization (ACO), genetic algorithm (GA), particle swarm optimization (PSO), grey wolf optimization (GWO), differential evolution (DE), genetic algorithm with aggressive mutation (GAAM), Non-dominated Sorting Genetic (NSGA) algorithm, and Culling algorithm (CA) [1, 15].

Some studies have attempted to employ GAs to choose the best features for classification. However, these studies are either implemented to target the improvement in classification and/or are not fully exploited in the context of the MDD patients detection and classification [24, 25]. Therefore, the current literature is only focused on improving the classification performance and thus unable to provide a good interpretability of various affected brain functionality of MDD patients along with good classification.

In this paper, we propose a classification solution for MDD patients along with interpretable results. For this purpose, we use the non-dominated sorting genetic algorithm (NSGA-II) to minimize the features for classification by selecting the optimal features. The optimal features are then utilized for a KNN as well as an SVM classifier. The least number of features are then correlated with classification results to improve the interpretability. By interpretability, we mean which brain functional activities are affected in an MDD patient. Note that this information may be crucial for a clinician for treatment and may provide more insight and knowledge of a MDD patient brain. Thus, improving the interpretability will not only help to detect MDD but may also help clinicians for better treatment of MDD patients. To improve interpretability, we have used GAs to select the minimum number of optimal features while maintaining a reasonable level of classification performance. Thus, NSGA-II allows us to select a minimum number of features for classification which eventually leads to correlating the classification results with the affected brain functionalities of MDD patients. To make the results reproducible, we open-sourced the code of this work at https://github.com/ehw-fit/eeg-mdd.

2 BACKGROUND

In this paper, our proposed approach consists of three stages; feature extraction from MDD patient's data, optimal feature selection, and classification of MDD patients using optimal selected features. Therefore, in this section, we review some algorithms and techniques that are prerequisites for our work.

2.1 Feature Extraction

In literature, several techniques based on time-domain, frequency-domain, and time-frequency-domain analysis of the signals are employed for feature extraction. Thus, feature extraction methods have been divided into three categories including time decomposition (TD), spectral analysis (SA) methods, and time-frequency analysis (TFA) methods [13]. We have initially investigated several feature extraction methods such as sample entropy, Hjorth parameter, Higuchi Fractal Dimension and power spectrum density

(PSD) using the Welch method. The Welch method is one of the widely used methods for the estimation of PSD. The weighted sum of the periodograms of the signal's overlapping windows is calculated using a non-parametric method. A windowed signal is used to generate overlapping segments and a discrete Fourier transform (DFT) is then determined for each segment. Finally, the Welch periodograms are produced by averaging the PSD for each segment. A mathematical formulation of the Welch method is shown in the following [9]:

$$\hat{I}_{xx}^{W}(w) = \frac{1}{P}\sum_{P=0}^{P-1}\hat{I}_{xx}^{P}(w) \tag{1}$$

where P shows the number of segments and $\hat{I}_{xx}^{P}(w)$ shows the periodogram per windowed segment, and $\hat{I}_{xx}^{W}(w)$ shows the average of $\hat{I}_{xx}^{P}(w)$ [9].

2.2 Evolutionary-based Feature Extraction

For EEG-based feature selection, several genetic algorithms such as CGA, GAAM, NSGA, and CA have been proposed in the literature [15]. The optimal feature selection methods for EEG are based on two criteria; the first is minimizing the features, and the second is maximizing the classification accuracy. Thus, this becomes a multi-criteria optimization problem. Since the NSGA method is dedicated to the multi-criterion optimization problem, therefore, we have selected the NSGA method for feature selection. The NSGA algorithm allows the finding of a subset of optimal features that leads to a higher classification capacity. Generally, in a decision process, several contradictory criteria are used. For an automatic decision, it is commonly useful to adopt one universal criterion to replace the contradictory criteria. If the choice is to be made automatically, it is appropriate to introduce one universal criterion that will replace the conflicting criteria. The idea of NSGA is an illustration of such a criterion [4]. NSGA is realized as a Holland algorithm with an explicit fitness function that operates around the concept of one individual dominating another. Those individuals who have worse values of all considered criteria compared to other individuals are considered dominated. In NSGA, for the evaluation of an individual, the recent population are sorted based on the domination principle and an appropriate index of Pareto front is assigned to each individual. Furthermore, from the subsequent Pareto fronts, individuals are determined based on the domination principle. In front 1, the individuals dominate by at least one criterion over the other fronts individuals. Similarly, in front 2 the individuals dominate by at least one criterion over the front 3 individuals, and so on. The fitness score given to an individual is represented by the front index. As a result, individuals from the front of the lower index are more appropriately suited to the optimization problem than those from the front of the higher index.

2.3 Standard Classifiers of MDDs

After the selection of optimal features, classification is the next stage. In literature, several machine learning algorithms have been proposed for EEG-based MDD classification, such as LR, SVM, and k-nearest neighbors (k-NN) etc., [6, 10]. Among all, LR performed well for a binary classification task, and requires low complexity. However, a high dimension of the input vector may lead to overfitting. In addition, it may also lead to the poor generalizability of the model. SVM is one of the alternative solutions for classification. Using SVM, the ideal boundary for categorizing data into two or more groups can be either linear or non-linear. However, its computational complexity is high for processing a large dataset. The k-NN algorithm has shown promising classification results and has been used in several EEG-based identifications of diseases, such as depression, anxiety, and epilepsy [6]. The k-NN classifies objects using a majority vote of the neighbors, assigning a case to the class that has the highest percentage of members among its k-nearest neighbors determined by a distance function. The algorithm includes three main aspects: a training set, similarity measure, and the parameter's size K. In this paper, we have used k-NN and SVM for classification purposes.

3 PROPOSED METHODOLOGY

In this work, we propose an automated methodology for feature selection. The overview of the methodology is shown in Fig. 1. The standard machine learning pipeline (red) is extended by evolutionary feature extraction. The candidate extractor is executed and the extracted features are fed into a given classification algorithm. The achieved sensitivity, specificity, and number of extracted features become desired design objectives to produce high-quality as well as simple classifiers. Accuracy is correlated to sensitivity and specificity and therefore, it is not included as an objective. The final classifiers should help the neuroscientists to interpret the results. Since the search algorithm is multiobjective. Therefore, we decided to use NSGA-II [4]. Since only high-quality feature-extractors are required, we also introduced a limit to the search.

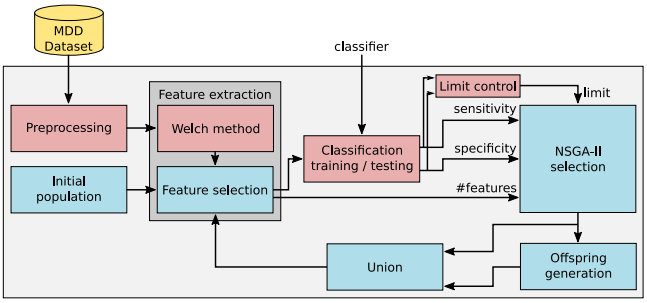

Figure 1: Overall methodology of the proposed searching algorithm.

3.1 Data Preparation

In EEG preprocessing, artifact-free EEG data is desirable to avoid subsequent erroneous analysis, ensuring that data genuinely represent the underlying neural activity. Therefore, in this study, the EEG preprocessing involved the correction of artifacts due to eye movements (horizontal and vertical), eye blinks, and muscular and heart activities. Moreover, the artifact corrections were performed with standard tools, including adaptive and surrogate filtering techniques, implemented in brain electrical source analysis (BESA) software [29]. In BESA, cleaning EEG data (artifact types: eyes

blinks, muscle activity, line-noise, heart activity, etc.) is based on a semi-automatic procedure; the technique has the name multiple source eye corrections (MSEC) [3]. According to this technique, the raw EEG data were used to estimate noise topographies. An appropriately selected head model (selected in BESA) and the noise topographies were used further to correct the artifacts. According to the procedure, an investigator needed to select the type of artifact (artifact types: eye blinks, muscle activity, line noise, heart activity, etc.) to be corrected. The selection allowed the software to mark the artifacts in the full EEG recording. The procedure is performed for all artifact types, including the artifacts due to eye blinks, eye movements, and muscular and heart activity. Hence, using BESA, the artifacts in the raw EEG were corrected for both the MDD patients and the healthy controls.

After preprocessing, the next step is feature extraction. The Power Spectrum Density (PSD) features are extracted from the brain's frontal, occipital, temporal and parietal regions. ANOVA is applied to them using MATLAB to statistically see the features' variance. The P-value for PSD for EEG bands is $p < 0.05$, showing the independence of the two groups statistically. That is why we choose PSD features for further analysis. The PSD is computed using the FFT with the Welch method and hamming window to estimate the power spectrum of the EEG time series with 2-seconds segments ($2 \times 250 = 500$ sample points), 50% overlapping (250 points) and the non-equidistant fast Fourier transform with 512 points. The PSD is estimated by the Welch method, which is an average of periodograms across time. When data is in a rectangular window, the periodogram consists of the non-overlapping blocks of data. The advantage of using Welch method is its capability to find the PSD of a signal by reducing the noise in the data. Also, it divides and takes the signal in small intervals by retrieving the maximum

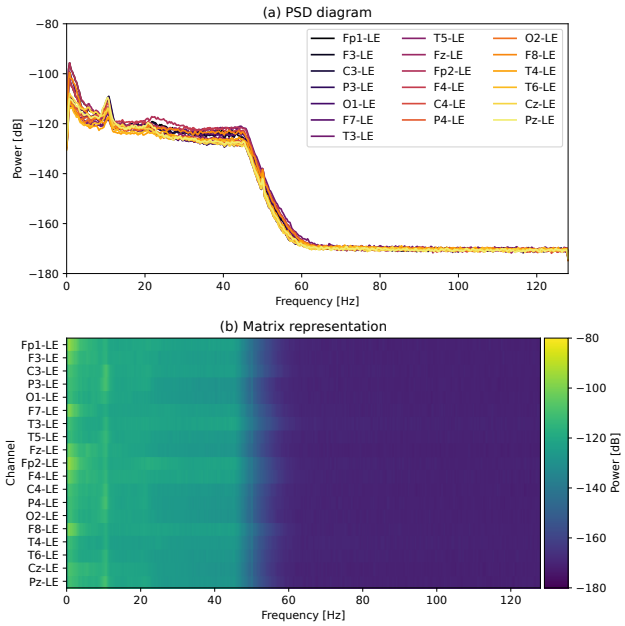

Figure 2: One subject from the MDD dataset represented as (a) PSD diagram (top), and (b) matrix heatmap (bottom).

information from non-linear EEG signal. The typical visualization of these data is a PSD diagram (Fig. 2a). However, for the ML feature, it is more effective to represent the data as a matrix heatmap (Fig. 2b).

3.2 Performance Parameters

In this paper, we have used accuracy, sensitivity, and specificity as quality performance parameters to evaluate the classification which takes the optimal selected EEG features as input. The performance parameters are briefly defined as follows:

Accuracy The percentage of MDD patients' EEG segments that are correctly classified is known as accuracy. Mathematically, it is calculated on the basis of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) and is defined as:

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

Sensitivity The accuracy rate of the positive samples is defined as sensitivity and is given as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

Specificity The accuracy rate of the negative samples is defined as specificity and is given as:

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

3.3 Evolutionary Feature Selection

As stated above, the goal of the evolutionary design program is to select appropriate channels (EEG electrodes) and frequencies from the power spectrum. The phenotype consists of a set of extractors. Each extractor is specified by a channel, ranges of frequencies, and function. The channel is specified by a name from the dataset. The ranges of frequencies are limited from 0 to half of the sampling frequency with a uniform step defined by Welch's method (e.g., 0.25 Hz). These ranges can be either unlimited or limited by so-called waveband (alpha, beta, etc.) — we proposed to limit them to 0, 4, 8, 12, 20, 30, and 50 Hz. We did not use frequencies above 50 Hz because the brain activity above 50 Hz is corrupted with noise [11]. The selected power features can be either used as they are (CP function), downsampled (*DS2/4/8* function), or aggregated (*AGG* function). Mathematically, the feature extractor is a function that transforms the power-spectrum matrix into a single vector of size c. An example of chromosome and phenotype is given in Fig. 3.

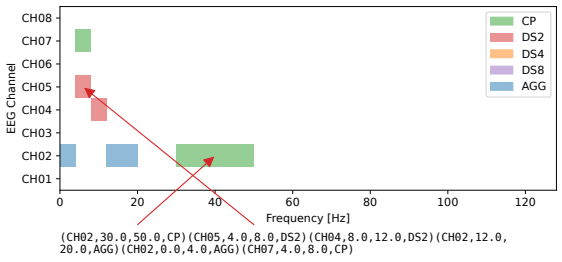

Figure 3: Example chromosome (bottom) and visualization of phenotype (top) of one candidate solution.

The first gene of the example selects a range of 20 Hz (with a step 0.25 Hz) as they are. That results in 80 features for each subject. The second gene extracts 8 features because 16 powers are downsampled by 2. The aggregation function produces four features (min, max, avg, amplitude) regardless of the frequency range.

Algorithm 1: Multiobjective feature selection

$P_0 \leftarrow \{i \in 0 \dots p : \text{randomIndividual}()\}$;
$limit \leftarrow 0$;
for $g \leftarrow 0$ **to** *gens* **do**
  $Q_g \leftarrow \{i \in 0 \dots q :$
   $\text{mutate}(\text{crossover}(\text{pickOne}(P_g), \text{pickOne}(P_g), m)\}$;
  $R_g \leftarrow P_g \cup Q_g$;
  evaluate(R_g) ; `/* train and test ML model */`
  `/* filter out weak individuals */`
  $R \leftarrow \{r \in R : \min(sens(r), spec(r)) \geq limit\}$;
  if $|R| > \alpha p$ **then**
    $limit \leftarrow limit + limit_{step}$; `/* increase limit */`
  end
  $P_{g+1} \leftarrow \text{nsgaSelect}(R, p)$;
end

Algorithm 1 shows the proposed search algorithm. The algorithm is inspired by the NSGA-II [4] algorithm. It is extended by an adaptive *quality limit*. First, a set of p random individuals is created. Then, for each generation, the following procedure is performed. First, a set of q offspring is constructed from the parent population. We propose to use uniform recombination at the level of genes. For each offspring, up to m values are mutated. The mutation can change the channel, the frequency range (in terms of waveband boundaries if needed), or the extraction function. The mutation can also drop one gene from the chromosome.

The entire population of feature extractors is evaluated. Each candidate extractor is applied to all N subjects, and a feature matrix X of size (N $timesc$) is constructed. This matrix X, together with a label vector y, enters the ML pipeline. Different quality metrics (accuracy, sensitivity, or specificity) and different training approaches (70:30 split, k-fold, etc.) can be used without loss of generality.

To avoid weak solutions, we propose to use an adaptive bound for the quality. The initial limit for sensitivity and specificity is set to 0. All candidate extractors leading to lower quality are removed from the R set. If more than $\alpha\%$ of the candidates have better quality, the quality limit is increased by a small value $limit_{step}$. According to our initial experiments, the introduction of this limit helped to fill the Pareto queue with more relevant solutions from a quality perspective. However, it was not confirmed that the absence of lower-quality solutions prevented the crossover operator from reducing the number of solutions found. Finally, a set of p non-dominated solutions is selected using Pareto filtering and distance crowding as proposed in NSGA-II [4].

4 EXPERIMENTAL SETUP

4.1 Dataset

This research uses a public dataset of MDD patients available from PLOS One [18]. This public dataset includes a sample of 32 MDD outpatients. The subjects were recruited based on the experiment design approved by the human ethics committee of the Hospital Universiti Sains Malaysia (HUSM), Kelantan, Malaysia. The study subjects signed the consent forms. The subjects were also briefed about the experiment protocol. The recruited MDD patients met the internationally recognized diagnostic criteria Diagnostic and Statistical Manual-IV (DSM-IV) for depression [2].

4.1.1 EEG data acquisition. An EEG cap with nineteen (19) electro-gel sensors was used to acquire EEG data. The electro-gel sensors required fewer adjustments than the hydro-sensors, facilitating longer recordings and enhanced patient care. The on-scalp placements of the EEG sensors followed the international 10–20 system [12]. According to the 10–20 system, the sensors are categorized into different regions, that is, the frontal region includes 7 electrodes namely Fp1, F3, F7, Fz, Fp2, F4, and F8, the central electrodes include C3, C4 and Cz, the parietal region includes P3, Pz and P4, the occipital involves O1 and O2, and the electrodes T3, T4, T5, T6 cover the left and right temporal regions [19]. An amplifier named Brain Master Discovery (Make: Brain Master, Model: Discovery 24e, Manufacturer: Brainmaster Technologies Inc.) was used to amplify the weak EEG signals from the sensors. Furthermore, the EEG data were digitized with 256 samples per second, band pass filtered from 0.1 to 70 Hz with an additional 50 Hz notch filter to suppress power line noise. The EEG data were recorded at pretreatment (before the start of medication). The EEG data were recorded during eyes open (EO) (5 minutes) conditions, while the study participants (MDD patients and healthy controls) were instructed to sit in a semi-recumbent position with minimal eye blinks and head movements.

4.2 Evolutionary Algorithm

Based on the initial experiments, the size of the parent population of NSGA-II algorithm was set to $p = 15$, and $q = 40$ offspring were generated in each generation. The change in the population size, except for extremely low or high values affected the convergence only. The found extractors were similar in terms of quality and size. The mutation rate m was up to 3. The evolutionary algorithm ran for 1,000 generations.

In the classification pipeline, we used the following machine-learning (ML) algorithms: k-NN using with Minkowski distance with p=2, and SVM due to its robustness. The ML implementations came from a well-established *sklearn* library. The accuracies (sensitivity and specificity) were achieved using the k-fold(5) algorithm. The dataset was divided into five folds. Then, in five iterations, every single fold served as test data and the rest as training data. This resulted in five accuracies that were averaged at the end.

To achieve statistically significant results, we performed four independent runs for each configuration. All presented experiments were executed on a computer equipped by AMD Ryzen 5 3600 6-Core Processor with 32 GB RAM.

5 RESULTS

5.1 Evolutionary Search

Fig. 4 shows the progress of each parameter during evolution. We can notice that at the beginning, the number of features is quite

high (more than 100), but very soon, we reach 20 features. The sensitivity and specificity follow the same trends at the beginning. However, the specificities higher than sensitivities occur at the end of the search. The limit (same shared value for both parameters) fastly reaches the value above 0.85 and keeps almost constant.

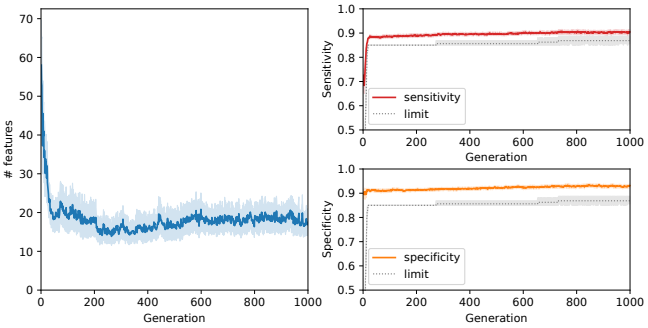

Figure 4: Average number of features, sensitivity, and specificity during the search. The light boxes show a confidence interval of 0.95 for all runs.

Table 1 summarizes the evolution parameters related to the selected classifier. The time spent in the search for 1,000 generations is related to the complexity of the training algorithm for particular models. The fastest training was achieved by the SVM algorithm. SVM and k-NN give comparable results; however, the number of features used by the SVM classifier is 32, and the number of features used by k-NN is 57. Since the MDD diagnosis is not a real-time application the diagnosis can be done offline. Our algorithm takes a maximum of 15 minutes, which is very realistic for this specific application. The complexity of the classifiers ($O(ndk)$ for k-NN and $O(n^3)$ for SVM) does not limit this application because of size of the dataset.

Table 1: Search algorithm results from 4 independent runs

Classifier	Time	Sensitivity			Specificity			# Features		
	avg. [s]	min	mean	max	min	mean	max	min	mean	max
k-NN	768.08	86.7%	89.4%	93.3%	87.1%	93.5%	96.7%	6	16.66	57
SVM	588.35	86.7%	92.0%	96.7%	87.1%	92.6%	96.7%	4	18.55	32

5.2 Analysis of Feature Extractors

Fig. 5 shows the final parent population (the Pareto optimal solutions and solutions close to the Pareto set) found from four independent runs for each classifier. In the left figure (sensitivity vs. specificity), we can see a regular grid. It is caused by a small dataset having 61 subjects only, and therefore, the granularity is not very smooth. Moreover, SVM and k-NN have six Pareto optimal solutions for sensitivity and specificity. The SVM classifier achieved 92% for both sensitivity and specificity (using 32 features - see Fig. 5 for more details). Such high accuracy is achieved, not only because of the quality of the classifier and selected features and the size of the dataset but also due to the number of folds in the k-fold algorithm.

As can be seen from Fig. 5, the number of selected features significantly affects the performance of the classifier. We evaluated the number of features needed to achieve sensitivity in Fig. 6. We

Figure 5: Tradeoff between design objectives. The Pareto optimal solutions for the pair of objectives are marked by a cross and joined by a line.

divided the found solutions into six classes by the number of features. For k-NN we can see that even with less than 10 features, the maximal sensitivity can be achieved. On the other hand, SVM classifiers need to have more than 20 features to reach the maximal sensitivity.

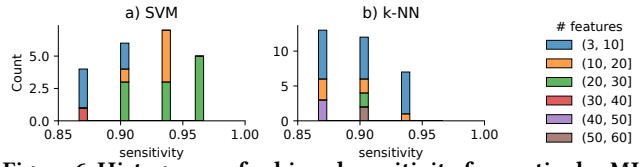

Figure 6: Histograms of achieved sensitivity for particular ML models depending on the number of the features extracted by the found extractors.

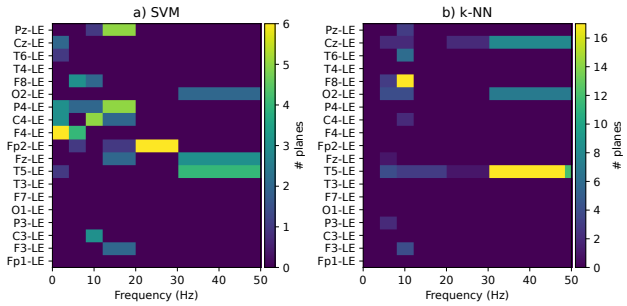

Figure 7: View how often each frequency range for different channels appeared in the set of Pareto-optimal feature extractors.

The frequency ranges for each channel are shown in Fig. 7 to show how frequently they appeared in the Pareto-optimal solutions. A detailed view is given in Fig. 8. It can be observed that the common frequencies for both classifiers are 4-12 Hz and 30-50 Hz; however, the SVM uses a broader frequency range. For SVM, the most frequent frequency ranges were 0-4 and 12-20 Hz. For k-NN classifier, the most frequent range was 30–50 Hz (gamma band). Since there is an overlap between k-NN and SVM extractors at range 30-50 Hz and channel T5, we found the gamma bands to be significant in MDD patients.

Fig. 9 shows an illustration of the histogram of extracted frequency planes in terms of electrode placement. Here for both k-NN and SVM classifiers, common activation is seen in the lateral T5

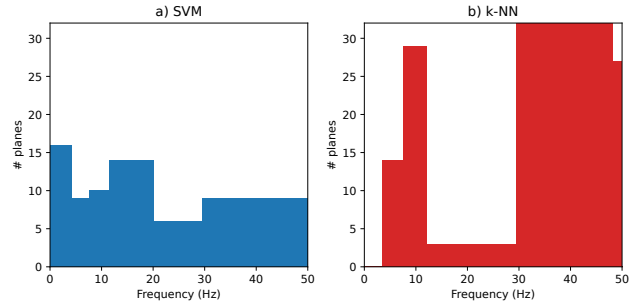

Figure 8: Histogram of frequencies used in found Pareto-optimal feature extractors.

electrode. In SVM features, C4 and P4 channels are seen in the final FEs. However, these features do not have the same frequency range (see Fig. 7, for example, F4 is used in 6 FEs for 0-4 Hz and in 4 FEs in range 4-8 Hz.)

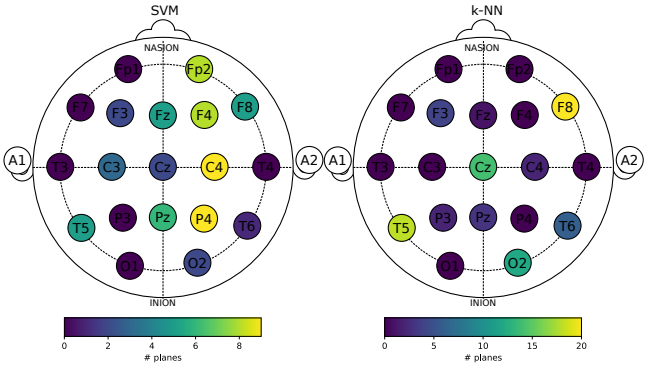

Figure 9: Histogram of channels usage in found Pareto-optimal feature extractors visualized on electrode placement map

5.3 Interpretation of the Results

In psychiatric clinics, diagnosing MDD is an iterative process based on the heuristic approach. In practice, sometimes patients with similar symptoms but different disorders than MDD are misdiagnosed and receive treatment as MDD patients. This conventional clinical practice can be improved by incorporating EEG data recordings and corresponding analyses that can help psychiatrists in providing an accurate diagnosis. This research work is a step in that direction.

Our primary finding in this research work is that the GA-based feature selection method using EEG modality is capable of diagnosing MDD patients. As seen from Fig. 9, there is an overlap in the left temporal region (T5 electrode) for both classifiers and also for the frontal region (F8 electrode). As MDD is gender biased, therefore for interpretability common to both genders, we select features from the left temporal region (T5 electrode) as they can classify MDD for both genders (males and females).

Existing literature also supports our findings. Vythilingam [30] found evidence for a lateralization effect, reporting smaller left temporal lobe volume in depression patients. Notably, the patient sample studied by these authors had the longest illness duration compared to other studies [30]. In this regard, the left-lateralized temporal lobe changes may reflect the progression of the disease over time or a distinct pathophysiological process that affects the risk of relapse, as well as brain volume abnormalities associated with MDD such as reduced BOLD responses in the temporal lobe. Such a condition will also result in changes in brain electrical activity in the left temporal lobe which can be detected by T5 electrode in EEG 10-20 system. Similarly, the hippocampus and amygdala are key brain structures of the medial temporal lobe, that are involved in cognitive and emotional processes, and are reported to be affected in MDD [22, 23]. The activity detected at T5 electrode location also incorporates the signals coming from these regions. Some other studies [5, 14, 26] also find that amygdala and left temporal lobe are affected in MDD patients as depicted in Fig. 10.

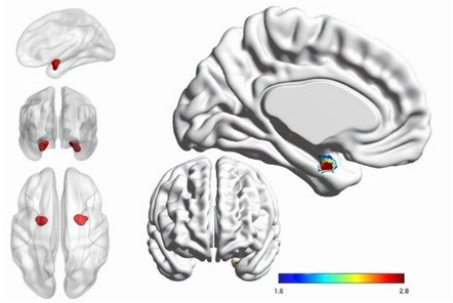

Figure 10: For MDD patients, the amygdala in left hemisphere is less active and less connected with other parts of the brain than normal control. [8].

To interpret the results, we selected one significant extractor for each classifier. As shown in Fig. 7, we select one of the Pareto-optimal extractors that employ some features with 60% occurrence in all extractors and with a maximal overlap of each to other. Fig. 11 shows two solutions, that is, solution #1 for the SVM and solution #2 for k-NN classifiers. It can be observed in Fig. 11a that solution #1 consists of channel T5-LE (electrode T5 with reference to Linked Ear LE) and its frequencies lie within the gamma band, (i.e., 30.0–49.75 Hz). A similar conclusion can be deduced from Fig. 11b and hence an overlap exists in the results of the two classifiers at channel T5-LE in frequencies range 30.0–49.75 Hz (gamma band). In both cases, the powers are aggregated to *avg*, *min*, *max* and *max-min* values. Thus, the power of gamma features extracted planes shows the temporal region's activation. Hence, we can conclude that for

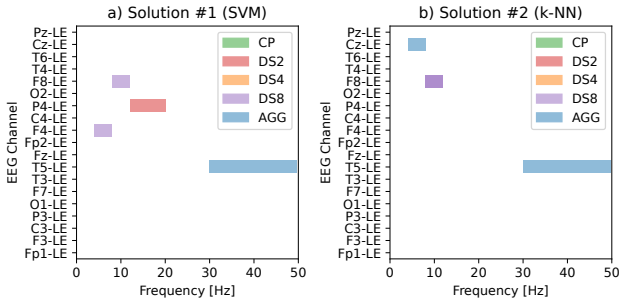

Figure 11: Visualization of the phenotypes of the selected feature extractors designed with (a) SVM and (b) k-NN classifiers

MDD patients, the dominant affected region of the brain is the left temporal region. Furthermore, in MDD patients' the most affected frequency band for both gender is the gamma band.

To generalize the results for interpretability, we cross-validate the two selected solutions for each classifier individually, and a combined version of the features extracted by the both selected solutions is also investigated for the accuracies of the SVM and k-NN classifiers in Table 2. In addition, the performance of both classifiers is also examined by using all features for all channels and frequencies (without any aggregation or downsampling). Since the k-fold algorithm provides k accuracies, they are statistically summarized in Table 2.

Table 2: Accuracy of cross-validation of the selected feature extractors, their combination, and all PSD features for all frequencies and channels for both classifiers evaluated using k-fold algorithm

Features extractor	SVM classifier			k-NN classifier		
	min	*mean*	*max*	*min*	*mean*	*max*
#1 (designed for SVM)	83.30%	**93.46%**	100.00%	66.70%	83.60%	91.70%
#2 (designed for k-NN)	50.00%	52.42%	58.30%	75.00%	**91.66%**	100.00%
#1 ∪ #2	50.00%	65.62%	83.30%	75.00%	88.46%	100.00%
all PSD powers	75.00%	88.46%	100.00%	75.00%	86.92%	100.00%

It can be observed from the results that a specialized extractor designed for a given ML model has the best quality (#1 for SVM and #2 for k-NN). Interestingly, such specialization (with downsampling) is better than a simple concatenation of both feature extractors. However, k-NN-based features cannot be used for SVM due to low sensitivity. On the other hand, both classifiers are not able to select the right features from all frequency ranges. It illustrates the fact that the performed task is non-trivial and the proposed algorithm significantly helps to improve the accuracy.

We also evaluated the results statistically using 100 independent k-fold classifications. Using the t-test, we found that the difference between extractors #1 and #2 is statistically significant (p-value < 0.05) for both classifiers. For k-NN classifier, FE #2 is significantly better than the others. However, for SVM classifier, the result (using all extractors) is strongly dependent on the splitting of the training and test data because of the variation of the classification accuracy. Therefore, it is challenging for an SVM classifier to determine the significance of these differences.

5.4 Comparison with manual approaches

A number of related methods are described in [6, 18, 24, 28]. The first three approaches [6, 18, 24] are based on machine learning and employ handcrafted features. Scalability is a problem with these approaches. The current methods use up to two types of features. In view of the large size feature vector for each type, adding more features results in a scalability issue. In contrast, the proposed method utilizes the GA method to address the scalability issue inherent in the existing state-of-the-art. In addition, the proposed GA-based method achieves higher accuracy despite using fewer features than the existing state-of-the-art method. A deep learning method proposed for MDD classification in [28] achieves 89.33% accuracy. Compared to an existing deep learning-based algorithm [28], the proposed method achieves a higher accuracy of 93.5%. The other advantage is interpretability; [6, 28] are not interpretable compared to a proposed method that provides interpretability. Overall, the proposed method has the edge over existing works regarding comprehensive interpretability and scalability. The soundness of the proposed method can be seen from the results as, for now, using just a few features, it outperforms the above mentioned state-of-the-art in terms of accuracy.

Table 3: Comparison with the State of the Art

Reference	Method	Accuracy
Duan et al. 2020 [6]	k-NN	82.45%
	SVM	86.15%
	CNN	93.50%
Mumtaz et al. 2016 [18]	logistic regression (LR)	87.50%
Saeedi et al. 2020 [24]	E-KNN	91.38%
Uyulan et al. 2020 [28]	CNN	89.33%
Proposed	SVM	93.46 %
	k-NN	91.66 %

6 CONCLUSION

This paper proposes an interpretable electroencephalogram (EEG)-based solution for classifying major depressive disorder (MDD) patients. This study gives EEG features as input data to the proposed GA method for MDD classification. The proposed GA-based method offers a new methodology that provides high efficiency (accuracy, sensitivity, and specificity) with fewer features. The results show that chromosomes from left lateral temporal areas can successfully identify MDD. A feature matrix was constructed involving frequency decomposition of EEG data based on power spectrum density (PSD). To improve the interpretability, the extractor selecting the features are designed via NSGA-II algorithm while minimizing number of features. The designed extractors are compared with each other to improve interpretability. From the results, it has been found that aggregated PSD of gamma bands extracted from the temporal regions can distinguish MDD from healthy control for both genders with an average sensitivity of 93.3%, specificity of 93.4% and accuracy of 93.5% using NSGA-II best solution for SVM and k-NN classifiers.

The source code of this work is available at https://github.com/ehw-fit/eeg-mdd.

ACKNOWLEDGMENTS

This work was supported by the Czech Science Foundation project 21-13001S.

REFERENCES

[1] Hesam Akbari, Muhammad Tariq Sadiq, et al. 2021. Depression recognition based on the reconstruction of phase space of EEG signals and geometrical features. *Applied Acoustics* 179 (8 2021), 108078.
[2] VA Arlington. 2013. Association, AP Diagnostic and Statistical Manual of Mental Disorders. *Am. Psychiatr. Assoc* 5 (2013), 612–613.
[3] Patrick Berg and Michael Scherg. 1994. A multiple source approach to the correction of eye artifacts. *Electroencephalography and clinical neurophysiology* 90, 3 (1994), 229–241.
[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Tr. Evolutionary Comp.* 6, 2 (2002), 182–197.
[5] Giuseppe Delvecchio, Philippe Fossati, et al. 2012. Common and distinct neural correlates of emotional processing in bipolar disorder and major depressive disorder: a voxel-based meta-analysis of functional magnetic resonance imaging studies. *European Neuropsychopharmacology* 22, 2 (2012), 100–113.

[6] Lijuan Duan, Huifeng Duan, et al. 2020. Machine Learning Approaches for MDD Detection and Emotion Decoding Using EEG Signals. *Frontiers in Human Neuroscience* 14 (9 2020).

[7] Catherine K Ettman, Salma M Abdalla, et al. 2020. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. *JAMA network open* 3 (2020). Issue 9.

[8] Westmead Institute for Medical Research. 2018. Looking inside the brain to distinguish bipolar from depression. https://www.westmeadinstitute.org.au/news-and-events/2018/looking-inside-the-brain-to-distinguish-bipolar-fr

[9] M Francis, Mihika Keran, Rachana Chetan, and B Krupa. 2021. EEG-Controlled Robot Navigation using Hjorth Parameters and Welch-PSD. *International Journal of Intelligent Engineering and Systems* 14 (8 2021), 231–240. Issue 4.

[10] Maria Camila Guerrero, Juan Sebastián Parada, and Helbert Eduardo Espitia. 2021. EEG signal analysis using classification techniques: Logistic regression, artificial neural networks, support vector machines, and convolutional neural networks. *Heliyon* 7 (6 2021), e07258. Issue 6.

[11] Stephan Hertweck, Desiée Weber, et al. 2019. Brain Activity in Virtual Reality: Assessing Signal Quality of High-Resolution EEG While Using Head-Mounted Displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 970–971.

[12] Herbert H Jasper. 1958. The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* 10 (1958), 370–375.

[13] Ashima Khosla, Padmavati Khandnor, and Trilok Chand. 2020. A comparative analysis of signal processing and classification methods for different applications based on EEG signals. *Biocybernetics and Biomedical Engineering* 40 (4 2020), 649–690. Issue 2.

[14] Mayuresh S Korgaonkar, May Erlinger, et al. 2019. Amygdala activation and connectivity to emotional processing distinguishes asymptomatic patients with bipolar disorders and unipolar depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 4, 4 (2019), 361–370.

[15] Krzysztof Lorenz and Izabela Rejer. 2015. Feature selection with NSGA and GAAM in EEG signals domain. In *2015 8th Int. Conf. on Human System Interaction (HSI)*. 94–98.

[16] Aamir Saeed Malik and Hafeez Ullah Amin. 2017. *Designing an EEG Experiment*. Elsevier. 1–30 pages.

[17] Aamir Saeed Malik and Wajid Mumtaz. 2019. *EEG-based experiment design for major depressive disorder: machine learning and psychiatric diagnosis*. Academic Press.

[18] Wajid Mumtaz, Aamir Saeed Malik, Mohd Azhar Mohd Yasin, Syed Saad Azhar Ali, and Likun Xia. 2016. EEG-based Diagnosis and Treatment Outcome Prediction for Major Depressive Disorder. (5 2016). https://doi.org/10.6084/m9.figshare.3385168.v1

[19] Wajid Mumtaz, Likun Xia, et al. 2017. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. *PLOS ONE* 12, 2 (02 2017), 1–30.

[20] OECD and European Union. 2018. *Health at a Glance: Europe 2018*. Paris/European Union, Brussels. 212 pages.

[21] Eric Pan and Jessica S. Rahman. 2021. Stress Recognition with EEG Signals Using Explainable Neural Networks and a Genetic Algorithm for Feature Selection. In *Neural Information Processing*. Springer International Publishing, 136–143.

[22] Francesca Pizzo, N Roehri, et al. 2019. Deep brain activities can be detected with magnetoencephalography. *Nature communications* 10, 1 (2019), 971.

[23] Kerry J Ressler. 2010. Amygdala activity, fear, and anxiety: modulation by stress. *Biological psychiatry* 67, 12 (2010), 1117–1119.

[24] Maryam Saeedi, Abdolkarim Saeedi, and Arash Maghsoudi. 2020. Major depressive disorder assessment via enhanced k-nearest neighbor method and EEG signals. *Physical and Engineering Sciences in Medicine* 43 (9 2020), 1007–1018.

[25] Aditi Sakalle, Pradeep Tomar, et al. 2022. Genetic Programming-Based Feature Selection for Emotion Classification Using EEG Signal. *Journal of Healthcare Engineering* 2022 (3 2022), 1–6.

[26] Anja Stuhrmann, Thomas Suslow, and Udo Dannlowski. 2011. Facial emotion processing in major depression: a systematic review of neuroimaging findings. *Biology of mood & anxiety disorders* 1, 1 (2011), 1–17.

[27] The Organisation for Economic Co-operation and Development (OECD). 2018. High costs of mental illness in Europe. https://www.oecd.org/health/mental-health-problems-costing-europe-heavily.htm

[28] Caglar Uyulan, Türker Tekin Ergüzel, et al. 2021. Major Depressive Disorder Classification Based on Different Convolutional Neural Network Models: Deep Learning Approach. *Clinical EEG and Neuroscience* 52 (1 2021), 38–51. Issue 1.

[29] Jayabal Velmurugan, Sanjib Sinha, et al. 2014. Magnetoencephalography recording and analysis. *Annals of Indian Academy of Neurology* 17, 5 (2014), 113.

[30] Meena Vythilingam, Eric Vermetten, et al. 2004. Hippocampal volume, memory, and cortisol status in major depressive disorder: effects of treatment. *Biological psychiatry* 56, 2 (2004), 101–112.

[31] World Health Organization. 2021. Depression and other common mental disorders: global health estimates.