Biomedical Signal Processing and Control 71 (2022) 103237

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Accurate classification of depression through optimized machine learning models on high-dimensional noisy data

Xingang Fang [a,*], Julia Klawohn [b], Alexander De Sabatino [a], Harsh Kundnani [a], Jonathan Ryan [b], Weikuan Yu [a], Greg Hajcak [b]

[a] Department of Computer Science, Florida State University, USA
[b] Department of Psychology, Florida State University, USA

ARTICLE INFO

Keywords:
EEG
ERP
Depression
MDD
Machine learning
Classification

ABSTRACT

Motivation: Depressive disorders are highly prevalent and impairing psychiatric conditions with neurocognitive abnormalities, including reduced event-related potential (ERP) measures of reward processing and emotional reactivity. Accurate classification of Major Depressive Disorder (MDD) based on ERP data could help improve our understanding of these alterations and propel novel diagnostic or screening measures. However, it has been particularly challenging due to the lack of generalization for noisy raw data with small sample sizes. We aim to improve classification performance for MDD using noisy ERP datasets using machine learning (ML) techniques. *Results:* We have developed two optimizations in our ML-based analysis of ERP datasets: effective feature extraction in the preprocessing of high-dimensional noisy data and enhanced classification through ensemble ML models. Together with a carefully designed validation strategy, our techniques provide a highly accurate method for MDD classification even for ERP data that are limited in sample size, inherently noisy and high-dimensional in nature. Our experimental results demonstrate that our ML optimizations achieve great accuracy and nearly perfect sensitivity simultaneously, particularly in classifying data samples unseen during the training process, compared to prior studies that perform regression-based classifications.
Supplementary information: A supplementary document on ERP data collection is available.

1. Introduction

Major depressive disorder (MDD), known colloquially as depression, is a frequent and serious mental disorder characterized by symptoms of depressed mood, loss of interest, and decreased energy. Depression is often chronic, recurrent, and comorbid. According to the World Health Organization (WHO), the proportion of the global population with depression in 2015 is estimated to be 4.4% [47]. It is also the leading cause of disease burden worldwide and was estimated to become the second most common cause of death and disability by 2020 [46].

Diagnostic determination of depressive disorders, such as major depressive disorder (MDD) or persistent depressive disorder (PDD), is usually performed using diagnostic interviews based on criteria from the Diagnostic and Statistical Manual of Mental Disorders [3]. Such interviews are time consuming and require specialized personnel. In addition, depressive disorders vary in symptomatology, and existing DSM subtypes have only limited predictive utility with regards to clinical course or therapy response. Moreover, the diagnostic criteria for MDD do not explain the etiopathogenesis of the disorder; depression is descriptive, not mechanistic. Increasingly, depression is understood to reflect abnormalities in the brain's reward system-evident in blunted neural response to reward and positive stimuli [17,38]. Thus, the quest for additional objective biomarkers to aid early diagnosis and prognosis of depressive disorders has recently attracted significant research interests. Some have explored the use of Machine Learning (ML) and artificial intelligence (AI) technologies because of their success and popularity in other domains. new platforms and modern technologies, new applications for machine learning are increasingly feasible [20,33].

Accurate classification of depression requires careful optimization and thorough training of Machine Learning models. To this end, we need to develop an ML model that has been trained by neural datasets collected from individuals clinically diagnosed with MDD and healthy controls (HC) from a previous study [25]. Then, based on the knowledge learned from the prior cases, the best models are selected to predict whether a new subject (with unseen neural measures) should be classified as having depression or not.

* Corresponding author.

https://doi.org/10.1016/j.bspc.2021.103237
Received 16 September 2020; Received in revised form 6 September 2021; Accepted 30 September 2021
Available online 22 October 2021

A wide variety of data sources can be used for MDD classification approaches, using either subjective methods such as questionnaires or description of symptoms, or objective assessment methods such as event-related potentials (ERP) or magnetic resonance imaging (MRI). ERP are collected as time-locked Electroencephalogram (EEG) activities. Both EEG [1,23,24,33,34] and MRI [20] have been employed as biomarkers in machine learning models for the diagnosis of MDD. For example, Mumtaz reported the application of the Support Vector Machine, Naive Bayes and Logistic Regression models in the diagnosis of MDD [34]. As a follow-up, the same team developed a machine learning framework that can leverage EEG-derived synchronization likelihood features and detect MDD patients with a relatively high accuracy [35]. With the same dataset, [31] developed four different ML-based classifiers for the detection of MDD patients based on the linear and nonlinear features of EEG signals. These recent publications demonstrate that ML models can be developed to improve diagnosis and classification of MDD patients using objective neural measures. The combination of multiple diagnostic methods may provide even better predictive performance.

Event-related potentials (ERPs) are direct measures of brain's neural responses to events, derived from the ongoing EEG. They have been shown to be robust measures of neurocognitive functions with excellent psychometric properties [22] and can relate to both individual differences in depressive symptoms and categorical clinical diagnoses of depression. Furthermore, ERP measures allow for analysis and quantification of neural processing of events with high temporal resolution at the scale of milliseconds. Moreover, they involve relatively low cost, and can be collected and analyzed relatively rapidly [29]. Unlike MRI, ERP data collection can be performed in diverse clinical settings and has very few contraindications. Thus, ERPs as measures of neurocognitive alterations in clinical depressive disorders are well-suited to support ML-based classification of MDD patients.

In the context of depression, two neurocognitive functional alterations have been studied using ERPs – reward insensitivity and impaired emotional reactivity [38]. Both dysfunctions have been put forward as mechanisms of anhedonia, a core symptom of depression [17]. ERP studies on reward dysfunction in depression focus on the reward positivity (RewP), an ERP evident when participants win money in simple guessing tasks. The RewP is maximal approximately 250 to 350 ms (ms) following feedback indicating monetary gains and is absent or reduced following losses [38], thus, the RewP is commonly measured as the difference between the ERP response to gains minus losses. The RewP has good psychometric properties [26], and relates to both behavioral [8] and fMRI measures of reward circuit function [6,11]. Critically, the RewP has been found to be reduced in individuals with current clinical depression [10,16,25,28] Furthermore, it was demonstrated that RewP improved sensitivity and positive predictive values in the classification of first-onset depressive disorders when used in conjunction with baseline depressive symptoms [36].

With regard to impairments in emotional reactivity in depression, the processing of emotionally evocative stimuli has been studied using the late positive potential (LPP), a stimulus-locked ERP component that is increased following the presentation of emotional content [14]. The increased LPP covaries with emotional arousal and is thought to reflect increased attention to motivationally salient stimulus content [16,41] and has been shown to possess good psychometric properties [32]. Previous work shows blunted neural response to emotional pictures, as indicated by smaller amplitude of the LPP in individuals with current depression [15,30,41–43].

Both LPP and RewP were recently assessed together in the same relatively large sample of depressed adults [25]. We found that both reduced RewP and LPP independently predicted depression status. In addition, the differentiation between the depressed and healthy groups was improved when both ERP measures were employed in combination. However, our prediction was mostly made with a regression model, and was not optimized on the predictive performance on unseen data. It is important to examine if other contemporary ML models or an ensemble of multiple ML models can provide further improvement on the prediction accuracy, particularly on data unseen by the model during training.

To this end, we have developed a framework that can optimize ML models for depression classification using ERPs. Our framework tackles the noisy nature of ERP measures and its impact on the accuracy of unseen data through two methods: engineered feature extraction and principal component analysis for dimension reduction [4,5] of noisy data. In addition, we have trained a total of seven ML models including the Random Forest [7] and ExtraTree(Extremely Randomized Trees) [21] models, which are well-known to address overfitting problem, to identify a few selected models for their predictive efficacy, and created a stacking ensemble ML models based on the base ML models. Some ensemble base models and the stacking ensemble model carry built-in noise reduction capability. Finally, the framework has employed a validation strategy through a combination of cross-validation and independent holdout testing techniques to minimize the overfitting issue on unseen data.

Our experimental results demonstrate that our ML optimizations achieve great accuracy and nearly perfect sensitivity simultaneously, particularly in classifying data samples unseen during the training process, compared to prior studies that perform regression-based classifications.

2. Methods

All EEG and clinical data used for the current study stem from a dataset previously examined by means of classic ERP quantification and regression-based analysis [25] and were re-analyzed here using ML techniques. The details of ERP data collection and processing are available from the supplementary document.

EEG data of adequate quality were available for 81 MDD and 43 HC participants for the reward task (RewP), 80 MDD and 42 HC participants for the picture viewing task (LPP), and 78 MDD and 40 HC participants for analyses that combined data from both tasks.

Starting from these EEG clinical data, we have formulated a framework that can develop optimized ML models for depression classification using ERPs, as shown in Fig. 1. Using both RewP and LPP datasets, our framework for the optimization of ML models consists of two main phases, *data pre-processing* and *model development*, which are performed in an iterative manner.

First of all, we formalize the RewP and LPP datasets into a standardized structure conformant to all the ML models, which is shown in the first *data pre-processing* phase in Fig. 1. For our development of optimized ML models on a small set of RewP and LPP samples, the lack of generalization is a major challenge. Specifically, because the number of samples is limited, the random noise from samples cannot cancel each other and will be assimilated by the model. Therefore, the learned model will perform poorly on future unseen data, a problem also known as overfitting or high variance issue. To cope with the noisy nature of the ERP measures and its effect on the accuracy of unseen data, we have employed engineered feature extraction and principal component analysis for dimension reduction of noisy data. Because the dimension reduction methods require hyperparameter tuning together with downstream ML models, it is performed in the second *model development* stage inside the iteration whereas the feature extraction is performed in the first *pre-processing* stage. Furthermore, we have trained a total of seven base ML models including Random Forest and ExtraTree and created a stacking ensemble ML models based on the base ML models. Finally, the validation strategy of the combination of cross-validation (CV) and holdout testing techniques was employed to minimize the overfitting issue on unseen data.

For each iteration, we evaluate a combination of features with various candidate models and model hyperparameters based on carefully selected classification metrics. The top candidate models exhibiting

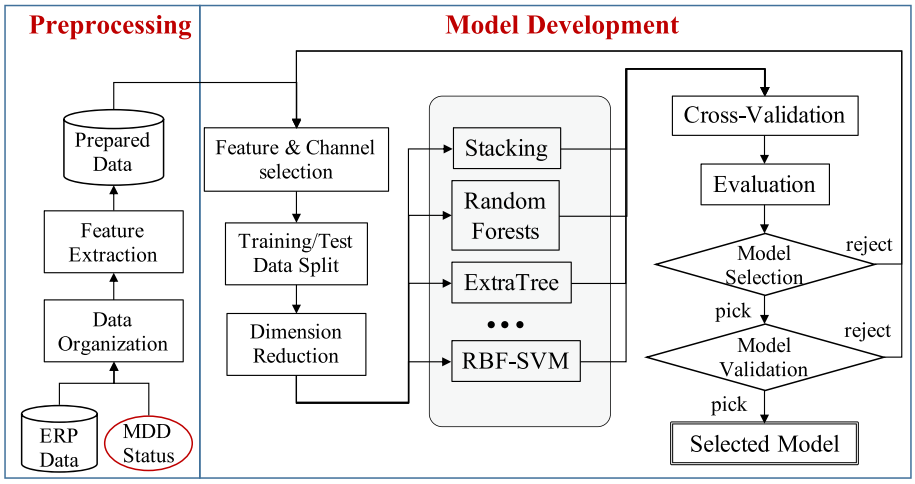

Fig. 1. Optimization of ML Models for MDD Classification.

best CV metrics are further tested on an independent holdout dataset to confirm that the achieved performance would reproduce on future unseen data. The results are leveraged in the next iteration for further evaluation of data pre-processing methods and ML models. In the rest of the section, we describe these phases in detail.

2.1. Data pre-processing

2.1.1. Data organization

Data Exploration: The first step of the *data pre-processing* was to explore and understand the data. The smallest units in the raw data were time-domain electrical potential value vectors collected from many EEG channels. The signals were collected in a certain period at a certain frequency, so the sizes of the vectors were fixed. For each participant, the signal vectors were collected from up to 31 EEG channels twice per channel for gain/loss (RewP) or positive/neutral (LPP) tasks. The diagnostic status of the participants in terms of a depression diagnosis was known. Through the exploratory analysis of the dataset, we confirmed that the signal vectors were collected without outliers or missing values. Missing channels due to artifact rejection were identified. The number of participants was at the scale of a hundred which may be relatively small for some models.

Labeling: General ML models require a dataset to be organized as a two-dimensional table of numeric values with row-based samples and column-based features. Classification models as employed in our research require a class label for each sample vector.

Because the goal was to identify potential participants with a depression diagnosis, by default, we set the "depressed" label as positive (1) and the "healthy" label as negative (0). This arrangement of class labels aligned with the purpose of this work to differentiate potential depressed participants from the healthy controls. This arrangement affects the statistical metrics which we will further elaborate in Section 2.2.2.

Organization Methods: Several methods of data organization were attempted. The reason for comparing these variations was to effectively address two major challenges of our dataset: 1. relatively small sample size; and 2. missing channels for some participants.

Small sample size is a common challenge in experimental clinical science because of the high recruitment effort and costs involved in clinical data collection. An intuitive way to organize our data was to prepare a single vector for each participant. It resulted in a dataset with the same number of samples as the number of participants (i.e., at the scale of around a hundred). The relatively small sample size might cause high variance (overfitting) issues and hurt the predictive performance on unseen data. We attempted several methods to address potential overfitting problems: employment of extracted features as invariants of the raw signal; dimension reduction before feeding to the ML model; intrinsic randomness (i.e. BAGGING ensemble models) for some models.

After the sample vectors were organized one-per-participant, the second challenge was to handle the variable subset of the numerous EEG channels collected from each participant. We needed to effectively organize the data into a uniformed format without either missing values or variable length vectors. The first approach, as a widely employed strategy, was to ask the domain experts to manually pick the best channel(s) that are collected for all participants. It has worked well in past works, but risks information loss from subjective human decision. Another more data-driven method is to simply ignore participants with missing channels or channels with missing participants. The latter is more common because participants were usually more important than channels. This family of methods can introduce more computation.

Our methods intended to generate samples one-per-participant of features based on manually picked channel(s). According to the past research, RewP is typically measured through channel Cz, FCz [16], and LPP through channel Pz [27], which are all on sagittal plane. Thus, the datasets with one or more empirically selected sagittal plane channels were prepared for comparison.

2.1.2. Feature extraction

For each EEG channel of a participant, ERPs included gain and loss conditions for the RewP dataset, and pleasant (positive) and neutral ERPs for the LPP data. The difference signals were calculated for both

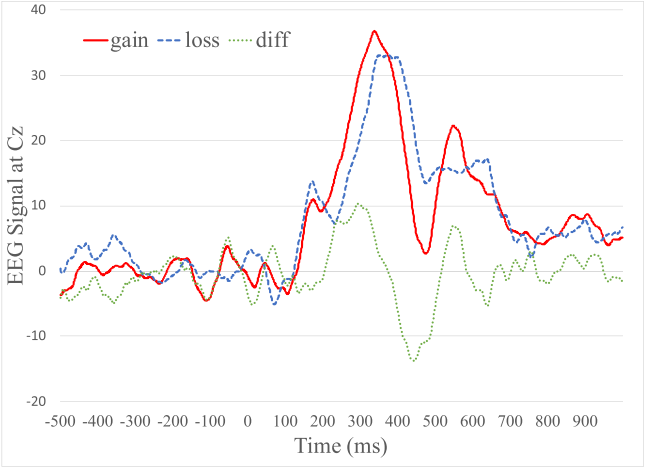

Fig. 2. Example RewP signals at channel Cz.

datasets. Fig. 2 shows a typical RewP ERP signal at one EEG channel for both gain/loss signals and their difference.

Next, we had the option to directly use ERPs as raw features (option 1) or perform feature extraction on the raw signals to generate more effective features and reduce noises. Without feature extraction, the raw signal dataset was noisy and may require a built-in noise reduction capability for the ML model to obtain good performance. Feature extraction was performed either manually during the *data pre-processing* (option 2) or by adding a dimension reduction model during the *model development* (option 3). All three methods were attempted, and the resulting model performances were compared.

Eight statistical features were extracted manually by applying statistical aggregation functions as listed in Table 1 to ERP signals of each channel. They are common time invariants extracted from time-series signals with minimal noise.

Another important family of feature extraction methods commonly employed in EEG datasets was the band power analysis. In our case, the data was transformed to frequency domain using either Fast Fourier Transformation or the Welch method and the average power in the delta, theta, alpha, beta, and gamma frequency ranges were collected as the band features.

In addition to manual feature extraction, three dimension reduction models known as the principal component analysis (PCA) [9], locally-linear embedding (LLE) [39] and Isometric mapping (Isomap) [40] were attempted. PCA was selected in model development for its fast computation and improved predictive performance in the preliminary model exploration.

A standardizing transformation was applied to the raw ERP signals before feeding them to the machine learning model. This would typically enhance the speed of the model convergence. After this standardization, the raw features were ready for the *model development*.

2.2. Model development

Model selection and optimization were guided by its performance in terms of the CV accuracy score on the training data. Fivefold CV was employed according to the preliminary results. Other important CV metrics such as precision, sensitivity/recall, specificity, etc. were considered in the final selection of the best candidate models. In the development, the average training metrics and the validation metrics, as well as their standard deviations, were examined to monitor the potential overfitting, underfitting, and outlier problems.

We chose a subset of available ML classification models according to the characteristics of our dataset. The ensemble models, especially the BAGGING (Bootstrap aggregating) models such as the Random Forest model [7] and ExtraTrees (Extreme Randomized Trees) model [21], were the focus of the study as they were known to reduce overfitting and increase the effective sample size by using random subsets of samples.

Support vector machine models [13] using either the linear or the radial basis function (RBF) kernels were selected for their capability of reducing overfitting. More models such as K-nearest neighbor [2], AdaBoost [18], gradient boosting trees [19,12], were also included in comparison for diversity.

A dimension reduction step mentioned in Section 2.1.2, though considered as a data pre-processing step, was closely integrated into the model development. When high-dimensional samples were fed to the model, a dimension reduction model such as the principal component analysis (PCA) model was optionally pipelined to the ML model.

An ensemble method known as "stacking" was proposed as they were known to be able to aggregate multiple weaker models and make a better model [45]. In this approach, multiple base models were employed to make out-of-fold predictions on the training data and their combined predictions were aggregated and served as new features for the consumption of a second level meta-model (usually a logistic regression model) to make the final prediction. The stacking ensemble model is a versatile approach to combine the power of multiple models and take advantage of all models. Stacking models were known to be robust when multiple uncorrelated ML models are included as base models. A robust ML model will provide more reproducible results.

Hyperparameters of top models were optimized using a hyperparameter-pipeline optimizer from the Scikit-Learn library [37].

Table 1
Statistical features from manual feature extraction.

Name	Description
Maximum	The biggest value
Minimum	The smallest value
Range	The difference between the maximum and minimum
Mean	The average of the signals
Standard deviation	The variance level of the signals
Skewness	The skewness of the major peaks in the signals
Kurtosis	The shape of the major peak in the signals
Signal to noise	The ratio of the mean to the standard deviation

2.2.1. Validation strategy

We followed the state-of-the-art model validation strategy as follows. All samples in the dataset were split into two portions, the training set and the holdout (test) set, in a certain ratio, 80% vs 20% ratio in our work. The training set was employed in the model selection while the holdout set was employed in the model testing step. In the model selection step, the training set was further split into folds and a method known as cross-validation (CV) was employed to obtain the desired score/metric representing the predictive performance of a candidate model. With a small sample size, small folds of three or five were favored over the more commonly used ten. We went for five-folds according to a preliminary result that favored five folds. The model with a better CV score would be favored in the development. The holdout set would only be employed to test the candidate models from the model selection to confirm that their good CV scores can be reproduced on unseen data. To avoid information leakage, the score/metric obtained from the holdout test should only be used to accept/reject candidate models.

Because the numbers of positive and negative labels were not even, both the train-test split and following CV split were performed in a stratified way so the positive and negative samples were evenly distributed in splits.

The statistical significance of the test metrics were tested using the Wilson score interval [44] at the 95% confidence level.

2.2.2. Predictive performance metrics

Choosing the right statistical metric is essential in the model development because it tells us which model is better. In our class label setup, the depressed label is the positive label while the healthy label is the negative label. For classification problems, there were several statistical metrics to examine: The accuracy metric represents the overall predictive performance which indicates how many predictions are correct. The precision metric indicates how many positive predictions were correct out of all positive predictions. The sensitivity, a.k.a. recall, metric indicates how many positive predictions were correct out of true positive cases. The specificity metric indicates how many negative predictions were correct out of all true negative cases. The area under curve (AUC) metric of the receiver operating characteristic (ROC) curve indicates the overall accuracy of both positive and negative predictions but will handle unbalanced problems, where class labels vastly differ, better than the accuracy metric.

In our specific case, the ROC AUC metric is not helpful as the class labels of healthy and depressed participants are relatively balanced at the ratio of around 2 to 1. We consider the sensitivity/recall metric most important in the diagnostic-type application, as a high sensitivity score indicates that the number of misdiagnosed depressed participants is minimal. Along the same lines, accurately predicting healthy participants (the negative class) was relatively less significant, so the specificity metric was less relevant. However, because the sensitivity metrics

from all top models were so high that the values were too close to compare. The near perfect sensitivity also lead to the fact that the accuracy, precision, ROC AUC, F-score, and κ statistic all correlated. Thus, the accuracy metric was chosen as the metric to lead model development.

3. Results

3.1. Constructed datasets

The dataset with only raw signals and the dataset with all 13 common EEG channels were excluded from further testing for their poor performances and high computation overheads in the preliminary exploration.

RewP and LPP datasets were constructed as combinations of raw signals and extracted features. Three groups of features were prepared: 1. Standardized raw signals; 2. Statistical invariant features; 3. Frequency Bands features. The standardized raw signal features were generated by normalization of raw ERP signals using a standard scalar. The raw signals collected in the gain and loss tasks for the RewP (positive and neutral tasks for LPP dataset) as well as their difference signals (*gain* $-$ *loss* or *positive* $-$ *neutral*) were horizontally concatenated to form wide vectors as raw signal features. Eight statistical features and five band features were extracted as described in Section 2.1.2. The features were combined to construct the base datasets for model development.

Starting from the base datasets which contains all EEG channels, a single or all channel(s) were selected to produce the derived datasets (Table 3) for the downstream model selection.

3.2. Selected models

The first round of model development was based on the RewP data. As mentioned in Section 2.2.1, we use 5-fold CV because of its comparable metrics and smaller standard deviation when compared to 3-fold CV. For dimension reduction, several methods were tested. Besides LLE and Isomap, PCA was chosen for its similar performance and less computation time.

All models as listed in Table 2 in Section 2.2 with default parameters were tested on datasets consisting of Cz, FCz and Fz channels. Runs with datasets 2 and 4 have optionally enabled PCA for dimension reduction. The training performance of top models is listed in the Table 4. K-nearest neighbor (KNN) models were eliminated from further testing after low accuracy scores of 0.52 and 0.44 were observed on the holdout test dataset. Furthermore, the XGB model was also excluded because of compatibility issues with pipeline libraries and its similar performance to peer ensemble tree models from the Scikit learn library.

3.3. Classification results

The Linear SVM, ET, RF and RBF SVM models with or without PCA were employed in the hyperparameter optimization. The grid search tool from SciKit learn library was employed. A comprehensive grid search in the hyperparameter space was performed on several datasets first to obtain a good understanding on the various hyperparameters on the predictive performances. The grids of hyperparameters on all datasets were thus chosen as described: For support vector machine (SVM) models, we set the search space of C to $\{0.1, 1, 10\}$ and the search space of gamma to $\{0.1, 1\}$. For ensemble tree models such as RF and ET, we set the number of sub-classifiers to $\{10, 50, 100\}$ when the dataset include high-dimensional raw features and to $\{5, 10, 20, 50\}$ otherwise. When a PCA model was employed, the search space of the number of components was set to $\{10, 50, 100\}$. The grid search in the hyperparameter space was performed on all 16 derived RewP datasets and resulted in 492 runs.

Table 3
Datasets, Feature Combinations and Selected Channels.

Dataset	Type	Feature combinations	Selected channel(s)
1	RewP	Statistical	Fz, FCz, Cz
2	RewP	Raw + Statistical Bands	Fz, FCz, Cz
3	RewP	Bands + Statistical	Fz, FCz, Cz
4	RewP	Raw + Bands + Statistical Raw	Fz, FCz, Cz
5	LPP	Bands + Statistical	Cz, Pz
6	LPP	Raw + Bands + Statistical	Cz, Pz
7	Both	Raw + Bands	Cz, Pz
8	Both	Raw + Bands + Statistical	Cz, Pz

Table 2
Machine learning models.

Model #	Detail
1	Linear support vector machine (Linear SVM)
2	Radial-basis function support vector machine (RBF SVM)
3	K-nearest neighbor (KNN)
4	Ada boost (Ada)
5	Extreme gradient boosting tree (XGB)
6	Extra tree (ET)
7	Random forest (RF)
8	Stacking ensemble model, models 1–7 as base models

Table 4
Preliminary training (CV) performance with RewP datasets.

Dataset	Model	Accuracy	Precision	Sensitivity	ROC AUC
1	RBF SVC	0.657	0.657	1.000	0.517
2	RF	0.677	0.670	1.000	0.575
2	ET	0.667	0.663	1.000	0.465
2	XGB	0.677	0.706	0.877	0.622
4	KNN	0.677	0.700	0.892	0.584

The top results are presented in Table 5. Because of the high sensitivity scores lead to strong correlations among the accuracy, precision, ROC AUC, F-score, and κ statistic, only the accuracy, precision and sensitivity scores are listed in the table. A pattern observed from the runs was the ineffectiveness of band features, as none of the top combinations involved these features. The entry #7 was rejected as the CV training accuracy was only 0.657, and after further analysis, in both runs, all class labels were predicted to be positive, meaning that these models turned out to be doing no classification at all. A group (group #1) of combinations of statistical features and simple models that carried less internal parameters such as SVM models (#4) and ensemble models with small number of base estimators (#1) exhibited excellent performance. Most top combinations (group #2) (all entries except #1, #4, and #7) consisted of three key components: 1. BAGGING ensemble tree models (ET, RF) with relatively large numbers of base estimators (50 or 100); 2. PCA to reduce the dimension to 100; 3. Raw signals together with statistical features. In terms of the channel(s), either all three sagittal plane channels together or FCz and Cz provided high predictive performance. The preliminary exploration of the data and results of other runs showed a high correlation among these three sagittal plane channels, so it was not surprising to see similar performance among the runs on various channel combinations. We also observed that all models exhibited high sensitivity performance in cross-validation tests except #1 with small numbers of estimators. This suggests that more complex ensemble tree models are preferred for high sensitivity.

The final test of the model performance on the holdout test dataset were performed on the combination #4 representing group 1 and #6 representing group 2. The results were listed in Table 6. Both the accuracy and sensitivity metrics were comparable to the training results. The 95% confidence intervals were reported and confirmed the significance of the results. Thus these models do not suffer from high variance (overfitting) issues. They are expected to work well with future unseen data.

As the results of the original traditionally analyzed study, [25]

Table 5
Training performance with RewP datasets.

					Train			Test		
Run	Dataset	Channels	Model	Dim Red	Accuracy	Precision	Sensitivity	Accuracy	Precision	Sensitivity
1	1	FCz	RF-5	None	0.960	0.955	0.985	0.697	0.732	0.862
2	2	all	ET-100	PCA-100	1.000	1.000	1.000	0.687	0.678	1.000
3	2	Cz	RF-50	PCA-100	1.000	1.000	1.000	0.687	0.685	0.969
4	1	FCz	RBF SVM	None	1.000	1.000	1.000	0.677	0.671	1.000
5	2	all	ET-50	PCA-100	1.000	1.000	1.000	0.677	0.671	1.000
6	2	all	RF-100	PCA-100	1.000	1.000	1.000	0.667	0.667	0.984
7	1	all	RBF SVM	None	0.657	0.657	1.000	0.657	0.657	1.000

Table 6
Test performance with RewP datasets.

Run	Model	Accuracy	Sensitivity
4	RBF SVM	0.640 ± 0.04	1.000 ± 0.00
6	RF	0.640 ± 0.04	0.938 ± 0.01

showed that the incorporation of LPP data with RewP data into the group status prediction enhanced the predicative performance, we tested the same set of models on LPP data only and the datasets with both RewP and LPP data. The Cz and Pz channels were selected in the comparison according to prior domain knowledge. The top CV results were listed in Table 7. The trend was similar to that of the RewP datasets. The RBF kernel SVM model, stacked model, and ensemble tree models performed best among all models. It was interesting to see that the stacking model of all base models (entry 4 in Table 7) exhibited comparable performance. The predictive performance was also similar.

Similarly, the performance for Entries #1, #2 and #4 is confirmed by the final validation on the holdout test datasets to represent the two groups of models (shown in Table 8) on the combined RewP + LPP datasets. The accuracy scores were slightly different from that of the RewP only dataset albeit overall similar (around one more or less participant was correctly predicted). The 95% confidence intervals confirmed the significance of the results.

In our current study, the incorporation of the LPP datasets in addition to the RewP datasets was not affecting the predictive performance by much. This observation is in contrast with the prior work [25], where by integrating both RewP and LPP measures the accuracy of diagnostic status classification improved from 53% to up to 66% with a linear regression model. One possible reason for the relative non-relevance of the additional data in the current study could be that the machine learning models with larger internal complexity were able to learn enough important information from the RewP data alone. From the perspective of the ML models, the LPP dataset may have contained the same or related information as the RewP data, so the predictive performance was not improved by integrating both data sources. In contrast, for simpler linear regression family of models employed in the referred paper, the simpler model might have leveraged additional variance in the LPP to improve predictions from the RewP. Another difference compared to the linear regression-based model is that our ML models are optimized for better predictive performance on future data using models trained with past data. Thus, the results are not directly comparable.

4. Conclusion

In this research, we have extensively explored the combination space of datasets, models and model hyperparameters, and found two groups of combinations of features and models that provided excellent sensitivity metrics and high accuracy metrics.

The first group of combinations consisted of: 1. Statistical features; 2. A simple model like RBF SVM or RF with a small number of base estimators; 3. No dimension reduction. This group had the benefit of the smallest data and model sizes and fastest model executions. One drawback was the slightly lower sensitivity metrics. Because of the smaller number of parameters in the model, this group of models may not benefit from added data of same type in the future.

The second group of combinations consisted of: 1. Raw signals plus statistical features; 2. ensemble tree models with a relatively larger number of base estimators; 3. an optional PCA as dimension reduction to reduce the dimension to 100. This group exhibited perfect sensitivity metrics. It is also expected to benefit from added data of same type in the future because the intrinsically high complexity of the model. High sensitivity indicates the high confidence in the detection of depressed individuals, which is a valuable characteristic of a screening tool. Both group of models exhibited nearly perfect sensitivity with only a few (or no) participants in the test dataset misclassified as depressed.

The performance has been evaluated using the holdout(test) dataset constructed to minimize the possibility of overfitting. Our models are expected to work well with future unseen data. Accurate classification on unseen data is a major advantage of the ML models developed in the current study in comparison to the previous regression-based analyses of the same data. These traditional regression models would only be relevant for the specific datasets. In contrast, the ML models derived in the current study have been tested on the unseen holdout data and demonstrated to perform well on new datasets. However, our ML models could not gain much in their accuracy or sensitivity when the LPP data were included with RewP data. This is because the LPP data are collected from different stimuli and our ML models are not sensitive to ERP data of mixed compositions.

In conclusion, through a combination of an effective noise reduction method and ensemble ML models, we have developed a highly accurate method for MDD categorization. Our experimental results with extensive training and test datasets demonstrate that our optimized ML techniques achieve high accuracy and nearly perfect sensitivity - an innovation relative to prior studies that provides enhanced effectiveness of classifying neurocognitive alterations associated with MDD. Future studies could test this ML model in novel participant data, collected across multiple labs, to further examine sensitivity and classification accuracy.

Funding

Google LLC for partial funding of the original study. This work is also supported in part by the National Science Foundation awards 1744336 and 1763547. This work uses the NoleLand infrastructure that is funded by the U.S. National Science Foundation grant CNS-1822737. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CRediT authorship contribution statement

Xingang Fang: Methodology, Investigation, Software, Formal analysis, Validation, Writing - original draft, Visualization. **Julia Klawohn:** Data curation, Writing - review & editing. **Alexander De Sabatino:**

Table 7
Training performance with LPP and RewP + LPP datasets.

					Train			Test		
Run	Dataset	Channels	Model	Dim Red	Accuracy	Precision	Sensitivity	Accuracy	Precision	Sensitivity
1	6	Pz	Stacking	None	1.000	1.000	1.000	0.691	0.685	0.984
2	5	Cz	RBF SVM	None	0.992	0.988	1.000	0.670	0.678	0.952
3	6	Cz	ET	None	1.000	1.000	1.000	0.670	0.691	0.906
4	7	Pz	Stacking	None	1.000	1.000	1.000	0.670	0.674	0.968
5	8	Cz	RBF SVM	PCA-100	1.000	1.000	1.000	0.660	0.660	1.000

Table 8
Test performance with LPP and RewP + LPP datasets.

Run	Model	Accuracy	Sensitivity
1	Stacking	0.640 ± 0.04	0.938 ± 0.02
2	ET	0.640 ± 0.04	0.938 ± 0.02
4	Stacking	0.625 ± 0.04	0.938 ± 0.02

Investigation, Software, Writing - review & editing. **Harsh Kundnani:** Investigation. **Jonathan Ryan:** Writing - review & editing. **Weikuan Yu:** Conceptualization, Writing - review & editing, Supervision, Project administration. **Greg Hajcak:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Dr. Kristen Schmidt for clinical assessments, Alec Bruchnak and Nicholas Santopetro for data collection.

Appendix A. ERP data collection procedure

A.1. Participants

Participants were recruited from the local community of Florida State University (FSU). Participants were labeled as MDD positive if they met standard diagnostic criteria for a current mood disorder (major depressive episode or persistent depressive disorder) in the past two weeks. Exclusion criteria for the MDD group were the presence of a lifetime diagnosis of a bipolar or psychotic disorder, or a current substance or alcohol use disorder. Participants were labeled as healthy control if they had never met diagnostic criteria for a mood disorder and did not currently meet criteria for any other psychiatric disorder. Potential participants were invited to the lab for interview after they passed a SCID-based screening administered over the telephone. Groups were equated for age, gender, and level of education. Participants were informed about the purpose and procedural details before the experiments and provided informed written consent. The study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki and approved by the Florida State University Institutional Review Board. The final sample included 83 MDD individuals and 45 healthy control participants (HC). EEG data of adequate quality were available for 81 MDD and 43 HC participants for the reward task, 80 MDD and 42 HC participants for the picture viewing task, and 78 MDD and 40 HC participants for analyses that combined data from both tasks.

A.2. Measures

Presence of current and past mood disorders was assessed in all participants with the Structured Clinical Interview for DSM-5 (SCID-5-RV) First et al. (2016) by two PhD level clinical psychologists. Other past and present psychopathology was evaluated using the Mini International Neuropsychiatric Interview (M.I.N.I.) (Sheehan et al., 1997, 1998) updated for DSM-5 (version 7.0.2).

A.3. Electroencephalogram recording

The electroencephalogram (EEG) was recorded using an active electrode EEG-system (ActiCHamp, Brain Products GmbH) with 32 scalp electrodes positioned in accordance with the 10–20-system (ActiCAP, Brain Products GmbH). Electrode Cz served as the recording reference, a ground electrode was placed on the forehead, two further electrodes on both mastoids, and the electrooculogram (EOG) was recorded from four additional electrodes: two approximately 1 cm above and below the left eye, two at the outer canthi of both eyes. Continuous EEG signals were recorded at a sampling rate of 1000 Hz using a bandpass recording filter of 0.01 to 100 Hz.

A.4. EEG tasks

For the collection of RewP data, the Doors task was administered using the Presentation software (Neurobehavioral Systems, Albany, California). It consisted of three blocks of 20 trials, each trail began with the presentation of two identical images of doors. Participants were instructed to select the left or right door. They were informed that they could either win $0.50 or lose $0.25 on each trial. The images of the doors were presented until participants made a selection. A fixation cross was then displayed for 1000 ms, followed by a feedback stimulus presented for 2000 ms. An upward green arrow or a downward red arrow was displayed to indicate the gain or loss, respectively. Another fixation cross was presented for 1500 ms, followed by the prompt "Click for next round" to let the participant enter the next trial. In the 60 trials for each participant, 30 gain and loss feedback stimuli were presented in a pseudo-random order.

For the collection of LPP data, we utilized a picture viewing task with 60 pictures selected from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008), including 30 pleasant images (e.g. erotic and affiliative images) and 30 neutral images (e.g. objects, humans with

neutral facial expression [1]). Normative ratings indicated that the 30 pleasant images were rated as more pleasant (valence M = 6.76, SD = 0.34) than the 30 neutral images (valence M = 5.36, SD = 0.53). All pictures were presented in random order across three sets of 20 trials. Each trial began with the display of a fixation cross for a random duration of 500 to 900 ms followed by pictures for 1500 ms, spanning approximately 15 to 20 degrees of visual angle. After picture offset, a blank screen was presented for a period of 500–900 ms. Participants were instructed to focus on the screen and view the pictures.

A.5. RewP and LPP Raw Dataset Construction

Both raw EEG datasets were processed using Brain Vision Analyzer, Version 2.1 (Brain Products, Gilching, Germany) to extract the RewP and LPP measures. Data were referenced to the average of the mastoid electrodes. A bandpass filter from 0.01 to 30 Hz was applied.

For the Doors task (RewP data), feedback-locked epochs were extracted with a duration of 1500 ms, starting 500 ms before feedback onset. Data were corrected for eye movement artifacts using the algorithm developed by Gratton & Coles (1983). Segments that contained voltage steps > 50 mV between sample points, a voltage difference of 175 mV within a 400 ms interval, or a maximum voltage difference of < 0.5 mV within 100 ms intervals were automatically rejected for individual channels. Additional artifacts were identified and removed based on visual inspection. Baseline-correction was applied using the 200 ms pre-stimulus interval as baseline. Feedback-locked ERPs were averaged separately for gains and losses and exported for ML analysis using the complete data segment for all channels.

For the picture-viewing task (LPP data), epochs from 200 ms before until 1200 ms after picture onset were extracted. Processing phases were identical to those described above with the exception that stimulus-locked averages were calculated separately for pleasant and neutral images, and data was exported for all channels using the whole segment.

A total of 118 participants had both RewP and LPP data collected, among which 78 participants had MDD and 40 participants were healthy controls. The Cz and FCz channels were employed in further analyses of the RewP data, whereas the Pz and Cz channels were employed for analyses of the LPP data.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.bspc.2021.103237.

References

[1] M. Ahmadlou, H. Adeli, A. Adeli, Fractality analysis of frontal brain in major depressive disorder, International Journal of Psychophysiology 85 (2) (2012) 206–211.

[2] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.

[3] American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (dsm-5).

[4] F. Anowar, S. Sadaoui, B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne), Computer Science Review 40 (2021), 100378.

[5] Y. Bai, Z. Sun, B. Zeng, J. Long, L. Li, J.V. de Oliveira, C. Li, A comparison of dimension reduction techniques for support vector machine modeling of multi-parameter manufacturing quality prediction, Journal of Intelligent Manufacturing 30 (5) (2019) 2245–2256.

[6] M.P.I. Becker, A.M. Nitsch, W.H.R. Miltner, T. Straube, A single-trial estimation of the feedback-related negativity and its relation to bold responses in a time-estimation task, Journal of Neuroscience 34 (8) (2014) 3005–3012.

[7] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[8] J.N. Bress, G. Hajcak, Self-report and behavioral measures of reward sensitivity predict the feedback negativity, Psychophysiology 50 (7) (2013) 610–616.

[9] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (9) (2014) 2812–2831.

[10] C.J. Brush, P.J. Ehmann, G. Hajcak, E.A. Selby, B.L. Alderman, Using multilevel modeling to examine blunted neural responses to reward in major depression, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 3 (12) (2018) 1032–1039.

[11] J.M. Carlson, D. Foti, L.R. Mujica-Parodi, E. Harmon-Jones, G. Hajcak, Ventral striatal and medial prefrontal bold activation is correlated with reward-related electrocortical activity: A combined erp and fmri study, NeuroImage 57 (4) (2011) 1608–1616.

[12] T. Chen, 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.

[13] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[14] B.N. Cuthbert, H.T. Schupp, M.M. Bradley, N. Birbaumer, P.J. Lang, Brain potentials in affective picture processing: covariation with autonomic arousal and affective report, Biological Psychology 52 (2) (2000) 95–111.

[15] D. Foti, D.M. Olvet, D.N. Klein, G. Hajcak, Reduced electrocortical response to threatening faces in major depressive disorder, Depression and Anxiety 27 (9) (2010) 813–820.

[16] D. Foti, J.M. Carlson, C.L. Sauder, G.H. Proudfit, Reward dysfunction in major depression: Multimodal neuroimaging evidence for refining the melancholic phenotype, NeuroImage 101 (2014) 50–58.

[17] D. Foti, K.D. Novak, K.E. Hill, B.A. Oumeziane, Neurophysiological assessment of anhedonia in depression and schizophrenia, in: Neurobiology of abnormal emotion and motivated behaviors, Elsevier, 2018, pp. 242–256.

[18] Freund, Y., Schapire, R.E., et al. (1996). Experiments with a new boosting algorithm. In icml, volume 96, pages 148–156. Citeseer.

[19] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics (2001) 1189–1232.

[20] S. Gao, V.D. Calhoun, J. Sui, Machine learning in major depression: From classification to treatment outcome prediction, CNS Neuroscience & Therapeutics 24 (11) (2018) 1037–1052.

[21] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning 63 (1) (2006) 3–42.

[22] G. Hajcak, J. Klawohn, A. Meyer, The utility of event-related potentials in clinical psychology, Annual Review of Clinical Psychology 15 (2019) 71–95.

[23] B. Hosseinifard, M.H. Moradi, R. Rostami, Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from eeg signal, Computer Methods and Programs in Biomedicine 109 (3) (2013) 339–345.

[24] A. Khodayari-Rostamabad, J.P. Reilly, G.M. Hasey, H. de Bruin, D.J. MacCrimmon, A machine learning approach using eeg data to predict response to ssri treatment for major depressive disorder, Clinical Neurophysiology 124 (10) (2013) 1975–1985.

[25] J. Klawohn, K. Burani, A. Bruchnak, N. Santopetro, G. Hajcak, Reduced neural response to reward and pleasant pictures independently relate to depression, Psychological Medicine (2020) 1–9.

[26] A.R. Levinson, B.C. Speed, Z.P. Infantolino, G. Hajcak, Reliability of the electrocortical response to gains and losses in the doors task, Psychophysiology 54 (4) (2017) 601–607.

[27] A.R. Levinson, B.C. Speed, G. Hajcak, Neural response to pleasant pictures moderates prospective relationship between stress and depressive symptoms in adolescent girls, Journal of Clinical Child & Adolescent Psychology 48 (4) (2019) 643–655. PMID:29412004.

[28] W.-H. Liu, L.-Z. Wang, H.-R. Shang, Y. Shen, Z. Li, E.F. Cheung, R.C. Chan, The influence of anhedonia on feedback negativity in major depressive disorder, Neuropsychologia 53 (2014) 213–220.

[1] Pictures used in the current study (IAPS numbers); Pleasant images: 4599, 4604, 4606, 4607, 4608, 4611, 4623, 4624, 4641, 4643, 4650, 4651, 4652, 4656, 4658, 4659, 4660, 4664, 4668, 4670, 4676, 4680, 4683, 4687, 4689, 4693, 4694, 4695, 4697, 4698; Neutral images: 7025, 7150, 7491, 7175, 7055, 7010, 7034, 7002, 7185, 7161, 7041, 7000, 7004, 5471, 5740, 7547, 7500, 7081, 7061, 7546, 7490, 7096, 5390, 7504, 7095, 7510, 7165, 5726, 7489, 5750

[29] Lizio, R., Del Percio, C., Marzano, N., Soricelli, A., Yener, G.G., Başar, E., Mundi, C., De Rosa, S., Triggiani, A.I., Ferri, R., et al. (2016). Neurophysiological assessment of alzheimer's disease individuals by a single electroencephalographic marker. Journal of Alzheimer's disease, 49(1), 159–177.
[30] A. MacNamara, R. Kotov, G. Hajcak, Diagnostic and symptom-based predictors of emotional processing in generalized anxiety disorder and major depressive disorder: An event-related potential study, Cognitive therapy and research 40 (3) (2016) 275–289.
[31] S. Mahato, S. Paul, Detection of major depressive disorder using linear and non-linear features from eeg signals, Microsystem Technologies 25 (3) (2019) 1065–1076.
[32] T.P. Moran, A.A. Jendrusina, J.S. Moser, The psychometric properties of the late positive potential during emotion processing and regulation, Brain research 1516 (2013) 66–75.
[33] W. Mumtaz, A.S. Malik, M.A.M. Yasin, L. Xia, Review on eeg and erp predictive biomarkers for major depressive disorder, Biomedical Signal Processing and Control 22 (2015) 85–98.
[34] W. Mumtaz, L. Xia, S.S.A. Ali, M.A.M. Yasin, M. Hussain, A.S. Malik, Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd), Biomedical Signal Processing and Control 31 (2017) 108–115.
[35] W. Mumtaz, S.S.A. Ali, M.A.M. Yasin, A.S. Malik, A machine learning framework involving eeg-based functional connectivity to diagnose major depressive disorder (mdd), Medical & Biological Engineering & Computing 56 (2) (2018) 233–246.
[36] B.D. Nelson, G. Perlman, D.N. Klein, R. Kotov, G. Hajcak, Blunted neural response to rewards as a prospective predictor of the development of depression in adolescent girls, American Journal of Psychiatry 173 (12) (2016) 1223–1230.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[38] G.H. Proudfit, J.N. Bress, D. Foti, A. Kujawa, D.N. Klein, Depression and event-related potentials: Emotional disengagement and reward insensitivity, Current Opinion in Psychology 4 (2015) 110–113.
[39] Roweis, S.T. and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500), 2323–2326.
[40] Tenenbaum, J.B., De Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. science, 290(5500), 2319–2323.
[41] A. Weinberg, G. Perlman, R. Kotov, G. Hajcak, Depression and reduced neural response to emotional images: Distinction from anxiety, and importance of symptom dimensions and age of onset, Journal of Abnormal Psychology 125 (1) (2016) 26.
[42] D.J. Whalen, C.M. Sylvester, J.L. Luby, Depression and anxiety in preschoolers: A review of the past 7 years, Child and Adolescent Psychiatric Clinics 26 (3) (2017) 503–522.
[43] D.J. Whalen, K.E. Gilbert, D. Kelly, G. Hajcak, E.S. Kappenman, J.L. Luby, D. M. Barch, Preschool-onset major depressive disorder is characterized by electrocortical deficits in processing pleasant emotional pictures, Journal of Abnormal Child Psychology 48 (1) (2020) 91–108.
[44] E.B. Wilson, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association 22 (158) (1927) 209–212.
[45] D.H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.
[46] World Health Organization, The global burden of disease: 2004 update, World Health Organization, 2008.
[47] World Health Organization (2017). Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.