Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | JUNE 15 2023

Toward interpretability of machine learning methods for the classification of patients with major depressive disorder based on functional network measures

Special Collection: Nonlinear dynamics, synchronization and networks: Dedicated to Jürgen Kurths' 70th birthday

Andrey V. Andreev ; Semen A. Kurkin ; Drozdstoy Stoyanov ; Artem A. Badarin ; Rossitsa Paunova ; Alexander E. Hramov

Chaos 33, 063140 (2023)
https://doi.org/10.1063/5.0155567

Toward interpretability of machine learning methods for the classification of patients with major depressive disorder based on functional network measures

Cite as: Chaos 33, 063140 (2023); doi: 10.1063/5.0155567
Submitted: 21 April 2023 · Accepted: 24 May 2023 ·
Published Online: 15 June 2023

Andrey V. Andreev,[1] Semen A. Kurkin,[1] Drozdstoy Stoyanov,[2] Artem A. Badarin,[1] Rossitsa Paunova,[2] and Alexander E. Hramov[1,a)]

AFFILIATIONS

[1]Baltic Center for Neurotechnology and Artificial Intelligence, Immanuel Kant Baltic Federal University, 14, A. Nevskogo str., Kaliningrad 236016, Russia
[2]Department of Psychiatry and Medical Psychology, Research Institute, Medical University Plovdiv, 15A Vassil Aprilov Blvd., Plovdiv 4002, Bulgaria

Note: This paper is part of the Focus Issue on Nonlinear dynamics, synchronization and networks: Dedicated to Juergen Kurths' 70th birthday.

[a)]**Author to whom correspondence should be addressed:** aekhramov@kantiana.ru

ABSTRACT

We address the interpretability of the machine learning algorithm in the context of the relevant problem of discriminating between patients with major depressive disorder (MDD) and healthy controls using functional networks derived from resting-state functional magnetic resonance imaging data. We applied linear discriminant analysis (LDA) to the data from 35 MDD patients and 50 healthy controls to discriminate between the two groups utilizing functional networks' global measures as the features. We proposed the combined approach for feature selection based on statistical methods and the wrapper-type algorithm. This approach revealed that the groups are indistinguishable in the univariate feature space but become distinguishable in a three-dimensional feature space formed by the identified most important features: mean node strength, clustering coefficient, and the number of edges. LDA achieves the highest accuracy when considering the network with all connections or only the strongest ones. Our approach allowed us to analyze the separability of classes in the multidimensional feature space, which is critical for interpreting the results of machine learning models. We demonstrated that the parametric planes of the control and MDD groups rotate in the feature space with increasing the thresholding parameter and that their intersection increases with approaching the threshold of 0.45, for which classification accuracy is minimal. Overall, the combined approach for feature selection provides an effective and interpretable scenario for discriminating between MDD patients and healthy controls using measures of functional connectivity networks. This approach can be applied to other machine learning tasks to achieve high accuracy while ensuring the interpretability of the results.

Major depressive disorder (MDD) is a common and debilitating psychiatric illness that affects millions of people worldwide. Despite advancements in the understanding of its underlying mechanisms, the diagnosis and treatment of MDD remain a significant challenge. In recent years, functional connectivity analysis of brain activity has emerged as a promising approach for the identification and characterization of MDD. Additionally, machine learning (ML) algorithms have shown remarkable success in classifying patients with different psychiatric disorders based on functional neuroimaging data. One of the common approaches for the classification of patients with MDD is the application of machine learning algorithms directly to the functional connectivity matrices. However, the problem of explaining the obtained classification results remains partially unresolved.

In this study, we present a combined approach for the classification of patients with MDD using machine learning and complex networks theory. We consider global network measures (mean node strength, average shortest path length, number of edges, clustering coefficient, and small-world coefficient) as features because they are more robust to within-group variability than individual connections in the functional network. Our results demonstrate that a simple linear discriminant analysis achieves high accuracy (up to 83%) in two cases: when we use all network connections or when we use only the strongest ones. The highest contribution is made by mean node strength, clustering coefficient, and the number of edges.

I. INTRODUCTION

Analysis of functional networks formed in the brain is a powerful approach to studying brain functions in normal and pathological states.[1–5] Recently, functional networks have been increasingly used in the evaluation of patients with neurodegenerative disorders, particularly in combination with machine learning (ML) methods, due to the wide capabilities of ML in the analysis and classification of large datasets and the identification of complex patterns within them.[6–8] However, as ML models become increasingly sophisticated, there is a growing concern regarding their interpretability, particularly in the context of biomedical network data. Consequently, despite the success of ML, its further integration into the medical domain is possible only after the development and testing of strictly interpretable/explainable algorithms for preliminary diagnosis.[9–12] Thus, the development of interpretable ML-based classifiers is an urgent problem.[13–15]

The ML algorithm interpretability refers to the ability to understand and explain the reasons behind their predictions or decisions. In the biomedical domain, interpretability is crucial for gaining insights into the underlying mechanisms of diseases, identifying potential therapeutic targets, and aiding in clinical decision-making.[15] The interpretability of ML models for biomedical network data remains a significant challenge.[16–18] First, the relevance of the problem lies in the complexity of biomedical network data. The scale-free properties, modularity, and hierarchical structures in such networks are often observed at the same time, which makes them challenging to interpret using traditional statistical methods alone. ML algorithms offer the potential to extract meaningful patterns and relationships from these complex networks. However, their black-box nature, i.e., the lack of transparency in how they arrive at predictions, hinders the comprehension and trustworthiness of their results.[19] Second, the lack of a clear definition of basic concepts surrounding interpretability contributes to the difficulty in substantiating the problem. In the context of ML methods application for biomedical network data, interpretability can encompass several dimensions. These include feature importance or relevance, model structure understanding, and the ability to generate human-interpretable explanations for predictions. Feature importance refers to identifying the specific network components that significantly contribute to the model's predictions. Understanding the model's structure entails comprehending how the model produces the overall prediction.[20] Human-interpretable explanations involve presenting the model's output in a manner that can be easily understood and validated by domain experts.

Statistical analysis before the ML model training is considered as a straightforward approach to achieve ML interpretability, which may allow an understanding of what features distinguish a patient group from the control one.[21–23] Furthermore, the identified patterns could be used for interpretation in the medical field, such as precised diagnostic process, prediction of the outcome, etc. However, statistical methods do not always succeed in identifying significant features, especially in their complex combination, while feature selection approaches based on ML (such as filter-type methods, wrappers, random forest methods, minimum redundancy maximum relevance method, Shapley additive explanations, etc.) may show good results.[24–26] This is usually due to the large inter-subject variability and small sample sizes in the data obtained in neurophysiological biomedical experiments. We can assume that a more universal and informative approach here is the combination of statistical methods and conventional ML feature selection techniques.[27,28] The authors[27] statistically contrasted electroencephalographic spectral power between the classes in the representative group of subjects and then used these statistically identified features to train an artificial neural network to classify brain responses to specific visual stimulation. Statistical methods can also confirm and supplement the results of the feature selection algorithms.[28] All of this will help us to advance the solution to the interpretability problem of the developing ML model.

In this study, we explored the potential of the described approach for identifying major depressive disorder (MDD) in patients based on the analysis of resting-state functional magnetic resonance imaging (rs-fMRI) data. MDD is one of the most common psychiatric disorders in the world. It affects approximately 300×10^6 people globally and is associated with significant disability, morbidity, and mortality. The relevance for psychiatry of the problem of MDD diagnostics is due to the insufficiency, ambiguity, and subjectivity of conventional clinical assessments and subjective reports.[29] On the other hand, in psychiatry, there is a significant lack of biological proof in terms of the diagnostic process.[30–32] The diagnosis itself leans only on the subjectivity of the clinical assessment scales, patient report, and the experience of the physician. Recent studies have demonstrated the good diagnostic potential of ML approaches relying on the analysis of differences in the specific strengths of connections or other network measures in functional networks derived from fMRI data.[33–36] However, the problem of explaining the obtained classification results remains partially unresolved.[37]

Here, we analyze resting-state functional networks reconstructed from fMRI data. We employ the simple ML method, linear discriminant analysis (LDA), to differentiate patients with major depressive disorder (MDD) and healthy control subjects. To ensure the interpretability of the ML model, we adopt the combined approach involving feature selection based on significance and statistical analysis. We examine the peculiarities and nuances of this approach in the multidimensional feature space. One significant aspect of this work is that we consider global network measures (mean node strength, average shortest path length, number of edges, clustering coefficient, and small-world coefficient) as features because they are more robust to within-group variability than

individual connections in the functional network. This approach is supported by recent studies showing that network measures may provide a more accurate and reliable biomarker for MDD diagnosis.[26,38] In other words, we build the classifier based on comparing network topology characteristics. We believe that this approach enables us to identify and explain the specificities in the resting-state functional network topological features of MDD patients. Furthermore, we consider the essential issue of the principle choice of threshold for reconstructing a binary functional network from raw data following Ref. 26.

II. METHODS

Figure 1 illustrates a schematic representation of the research paradigm and overall structure of the research: the consequence of methods applied to the original data. A detailed description of the data and methods is presented below.

A. Experimental data

As experimental data, we used a set of 166 × 166 symmetric functional connectivity matrices calculated based on blood-oxygen-level dependent (BOLD) signals in 166 brain regions. The BOLD signal detected during rs-fMRI experiment reflects the changes in deoxyhemoglobin driven by spatially localized variations in brain blood flow and blood oxygenation, which are coupled to underlying neuronal activity by a process termed neurovascular coupling. We extracted the normalized fMRI volumes with the help of SPM 12 software and parcellated them into 166 regions according to the automated anatomical labeling atlas AAL3.[39] To estimate the connectivity between the brain regions, we calculated the Pearson correlation coefficients (absolute values)[40] for all pairs of parcel-averaged BOLD activities. The analyzed dataset contains connectivity matrices for 85 subjects: 50 healthy ones as a control group and 35 subjects with a major depressive disorder (MDD group).

Subjects from both groups were assessed by experienced psychiatrists using Mini International Neuropsychiatric Interview and Montgomery–Åsberg Depression Rating Scale (MADRS). Subjects having a previous history of comorbid psychiatric conditions, autoimmune diseases, neurological diseases, history of head trauma, or any metal implants incompatible with the MRI were excluded. All participants provided a written consent form complying with the Declaration of Helsinki. The study was approved by the Medical University of Plovdiv Ethical Committee (2/19 April 2018). The two groups of subjects did not differ significantly in terms of mean age, sex, and level of education distribution.

The MR scanning procedure was performed on a 3T MRI system (GE Discovery 750w, General Electric, Boston, MA, USA). The protocol included a high-resolution structural scan (Sag 3D T1) with a slice thickness of 1 mm, matrix 256 × 256, TR (relaxation time) 7.2 s, TE (echo time) 2.3 s, and flip angle 12°, FOV 24, 368 slices and resting-state functional scan–2D echo-planar imaging (EPI), with slice thickness 3 mm, matrix 64 × 64, TR 2000 ms, TE 30 ms, 36 slices, flip angle 90°, FOV 24, a total of 192 volumes. Before the EPI sequence, subjects were instructed to remain as still as possible with their eyes closed and not to think of anything in particular. The duration time of the resting-state functional scan was 6 min. MRI data were pre-processed in a typical way (see Sec. 2.1.3 in Ref. 36).

B. Network measures

Each functional connectivity matrix could be represented in the form of a network (graph). To analyze the network's structure and topology, we calculate the following global measures: mean node strength $\langle k\rangle$, average shortest path length $\langle L\rangle$, number of edges N_e, clustering coefficient C, and small-world coefficient σ.

Mean node strength is calculated as[41]

$$\langle k\rangle = \frac{1}{N}\sum_{i=1}^{N} k_i, \tag{1}$$

where k_i is the strength of ith node (the sum of weights of edges connected to the node) and N is the number of nodes in the graph.

Average shortest path length is calculated as[42]

$$\langle L\rangle = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij}}{N(N-1)}, \tag{2}$$

where L_{ij} is the shortest path between ith and jth nodes. Note that $L_{ii} = 0$ for $i = 1, \ldots, N$, so we exclude it from the calculation.

Clustering coefficient is the Watts–Strogatz clustering coefficient calculated as[43,44]

$$C = \frac{1}{N}\sum_{i=1}^{N} 2n_i/k_i(k_i - 1), \tag{3}$$

where n_i is the number of direct edges interconnecting the k_i nearest neighbors of node i.

Small-world coefficient is calculated as[45]

$$\sigma = \frac{C/C_r}{\langle L\rangle/\langle L_r\rangle}, \tag{4}$$

where C_r and $\langle L_r\rangle$ are the clustering coefficient and the average shortest path length for an Erdős–Rényi random graph with the same number of nodes and edges, respectively.

For measures calculation, we used open-source NetworkX package in Python.

C. Thresholding

The weights of edges (or connections) in the obtained functional networks are distributed in the range of 0–1. We introduce a threshold value Thr to remove the edges with weights $w < Thr$.[36] So, increasing Thr leads to leaving only strong connections in the network. In the process, some nodes can become disconnected (their strength is equal to 0); we remove these nodes from the network. Figure 2 shows the graph representation of the functional connectivity matrix of one healthy subject for different threshold values Thr = (a) 0, (b) 0.3, and (c) 0.6. The graphs were built using the Fruchterman–Reingold force-directed algorithm (realized in NetworkX package in Python), which simulates a force-directed representation of the network, treating edges as springs holding nodes close while treating nodes as repelling objects, sometimes called an anti-gravity force. The node size corresponds to its strength: the larger the node, the higher the strength. We change Thr in the

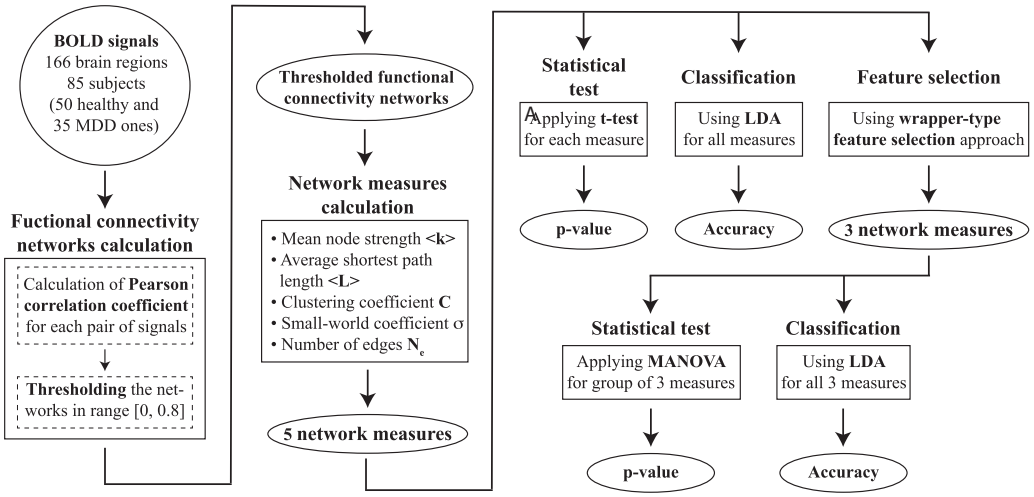

FIG. 1. Schematic representation of the overall structure. Rectangle frames depict steps of data analysis, and oval frames correspond to the input/output data for each step.

range [0, 0.8] because for $Thr > 0.8$, the graph is likely to become disconnected and splits into a number of small subgraphs.

D. Statistical analysis

To identify differences in network measures at the group level, we performed statistical testing both separately for each measure using the t-test and comprehensively for several network measures using multivariate analysis of variance (MANOVA). We used the Shapiro–Wilk test to check normality and Levene's test to check for equality of variance. We used the open-source statistical package Scipy in Python for statistical analysis.

E. Classification

To classify patients with MDD, we use Linear Discriminant Analysis (LDA), a supervised machine learning method that allows us to perform dimensionality reduction by projecting the input data to a linear subspace consisting of the directions which maximize the separation between classes.[46] We use a set of network measures as features.

First, we split each group of subjects (Control and MDD) into train and test subsets in the proportion of 60% by 40%. Then, we apply 100 random permutations for cross-validation, fit the LDA model with the train set, and test it with the test one by calculating the accuracy of the model,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

where FP (false positive) is indicated as the number of persons for a MDD group but erroneous for a Control group. FN (false negative) was indicated as the number of patients for a Control group but erroneous for a MDD group, and TP (true positive) and TN (true negative) are classified correctly.

As a result of applying LDA to the dataset, we get a separation hyper-plane defined as

$$\mathbf{w}^T\mathbf{x} + b = 0, \tag{6}$$

where $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$ is the vector of LDA coefficients, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is the vector of features, b is the constant LDA coefficient, and $\cdot^T$ is the operator of matrix transposition.

The distance from the ith subject to the separation hyper-plane is calculated as

$$D = \frac{\mathbf{w}^T\mathbf{x}_i + b}{||\mathbf{w}||}, \tag{7}$$

where $||\cdot||$ is a matrix norm.

For classification, we used the open-source scikit-learn package in Python.

F. Feature selection

To estimate the contribution of each feature, we propose a wrapper-type feature selection approach.[24] Initially, we use all N_f features for the LDA model training, then we remove one of them and use the rest $N_f - 1$ features. Then, we estimate the contribution of the ith feature by calculating the difference between the accuracies for the cases of N_f and $N_f - 1$ features,

$$\text{Contribution} = \text{Accuracy}(N_f) - \text{Accuracy}(N_f - 1). \tag{8}$$

We repeat the procedure for each discardable feature, so we have N_f repetitions. Then, we choose the feature with the lowest contribution and discard it. After that, we can repeat the whole procedure as many times as we need to leave a certain number of the most important features.

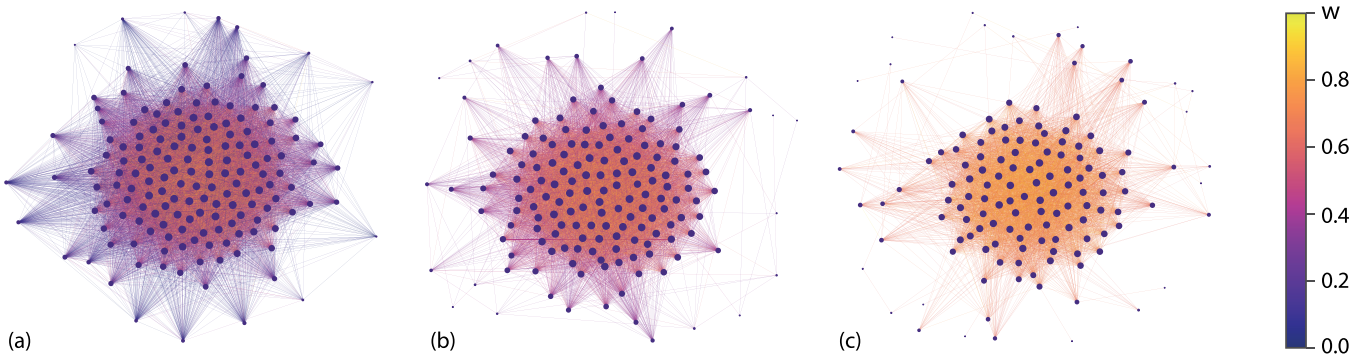

FIG. 2. The graph representation of the functional connectivity matrix of one healthy subject for different threshold values *Thr* = (a) 0, (b) 0.3, and (c) 0.6. The node size corresponds to its strength: the larger the node, the higher the strength. The edge color represents its weight *w*.

III. RESULTS

A. Analysis of network measures

For each *Thr*, we calculated the following network measures for all subjects: mean node strength $\langle k \rangle$ [Eq. (1)], average shortest path length $\langle L \rangle$ [Eq. (2)], number of edges N_e, clustering coefficient C [Eq. (3)], and small-world coefficient σ [Eq. (4)]. Figure 3 illustrates the dependencies of these measures on the threshold value *Thr* for both groups (MDD and Control). The mean node strength [Fig. 3(a)] and the number of edges [Fig. 3(c)] decrease monotonously with increasing *Thr*, with mean values of these measures being very close to each other for both groups, but the standard deviation for the MDD group is smaller than for the healthy one. On the contrary, the average shortest path length [Fig. 3(b)] increases linearly with increasing threshold, and it is equal for both groups for $Thr \in [0, 0.55]$, but for $Thr > 0.55$, $\langle L \rangle$ becomes higher for the MDD group. It means that the connectedness in the functional network is weaker for subjects with MDD for higher thresholds. The clustering coefficient [Fig. 3(d)] also linearly increases for $Thr \in [0, 0.65]$ for both groups, but from $Thr = 0.65$, it starts to decrease. One can see that for small threshold values, C is higher for the control group, but for $Thr = 0.6$, it becomes equal in both groups, and for higher threshold values, they change places. The dependencies of the small-world coefficient σ [Fig. 3(e)] are the most different from the other

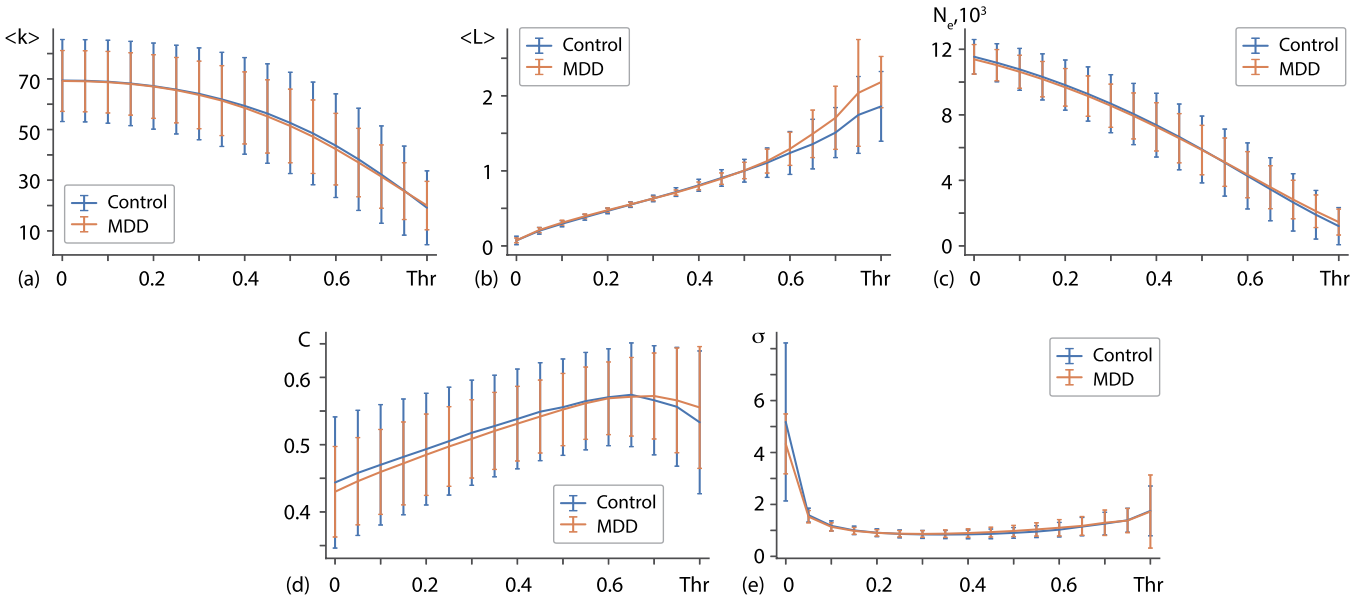

FIG. 3. The dependencies of the network measures on the threshold value *Thr* for Control (blue) and MDD (red) groups: (a) mean node strength $\langle k \rangle$, (b) average shortest path length $\langle L \rangle$, (c) number of edges N_e, (d) clustering coefficient C, and (e) small-world coefficient σ. Vertical lines correspond to the standard deviation of the measure inside the group.

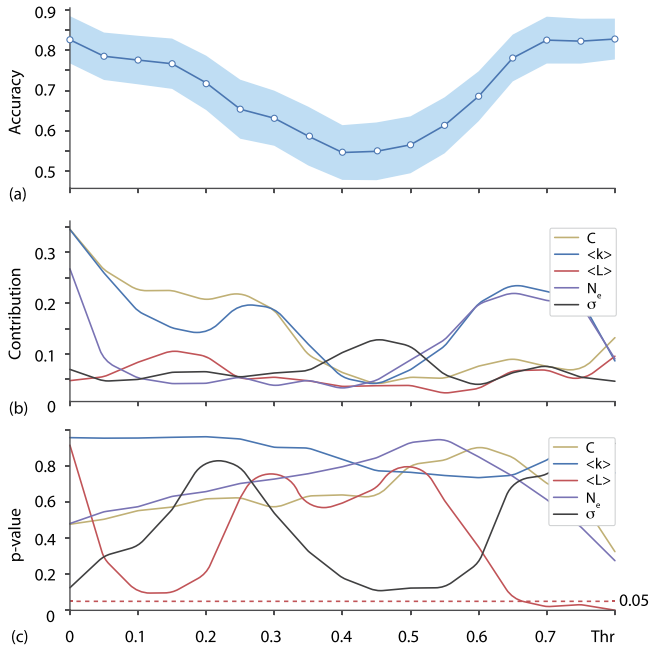

FIG. 4. The dependencies on the threshold value *Thr* of (a) the accuracy of classification (mean ± standard deviation due to random permutations) using all five features, (b) the contribution of each feature to classification outcome, and (c) the p-value of statistical separability of control and MDD groups using only one feature.

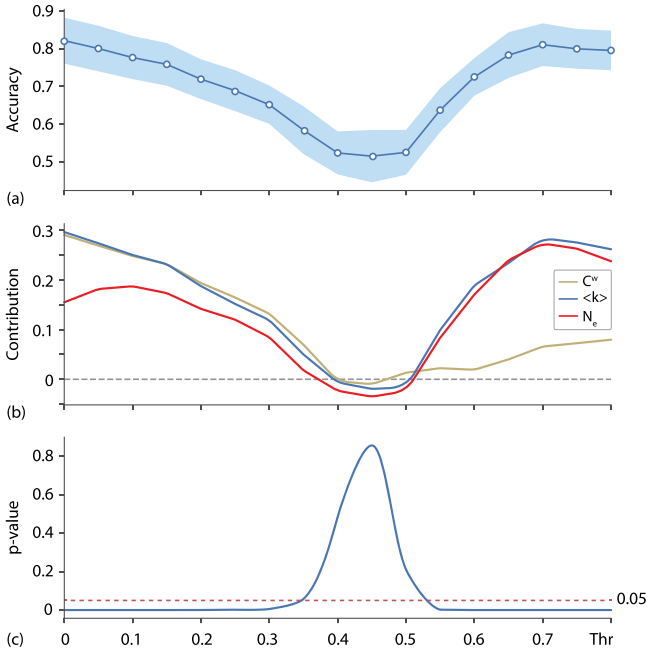

FIG. 5. The dependencies on the threshold value *Thr* of (a) the accuracy of classification (mean ± standard deviation due to random permutations) using only three features (mean node strength $\langle k \rangle$, number of edges N_e, and clustering coefficient C), (b) the contribution of each feature to classification outcome, and (c) the p-value of statistical separability of control and MDD groups using all three features together.

measures: first, they rapidly decrease for small *Thr* values, then vary in a small range of values, and σ is almost equal for both groups.

B. Classification and interpretability

We used the calculated network measures as features for binary classification of belonging the subject to one of two groups: Control or MDD group. Figure 4(a) shows how the accuracy of classification with LDA (mean ± standard deviation due to the random permutations) depends on threshold value *Thr*. As one can see, maximal accuracy is achieved for $Thr = 0$ (Accuracy $= 0.826 \pm 0.058$) and $Thr = 0.7$ (Accuracy $= 0.825 \pm 0.058$), while minimal Accuracy $= 0.547 \pm 0.067$ is achieved for $Thr = 0.4$. So, increasing *Thr* leads first to decreasing the accuracy to almost 0.5, but then it increases to the previous value. It means that for the best classification, we need to leave either all the connections in the networks or only the strongest ones with weights $w > 0.7$.

For the interpretability of classification results, we estimated the contribution [Eq. (8)] of each feature to the classification outcome using the wrapper-type approach described in Sec. II F. Figure 4(b) illustrates how much each feature contributes to the outcome accuracy (how much accuracy we lose by removing the feature) depending on *Thr*. We revealed that the most important features are the clustering coefficient C and the mean node strength $\langle k \rangle$ for $Thr < 0.38$, and the clustering coefficient C and the number of edges N_e for $Thr > 0.53$. For $Thr = 0$, the maximal contribution is made by C, N_e, and $\langle k \rangle$. With an increase in the threshold, the contribution of all of them decreases. For $Thr = 0.45$, these measures reach their minimal contribution, and with further threshold increasing, N_e and $\langle k \rangle$ become the most significant. Herewith, small-world coefficient σ being insignificant for most threshold values (contribution is around 0.05) becomes significant with a contribution of more than 0.1 for $0.38 < Thr < 0.53$ when the accuracy of classification is minimal.

Also, we estimate the statistical separability of each network measure for control and MDD groups. We consider each measure separately from others and apply a t-test. As a result, we obtain a p-value for each measure for the considered threshold value *Thr*. If the p-value < 0.05, the groups are considered to be statistically separable. Figure 4(c) illustrates the dependencies of the p-value for all measures on *Thr*. It shows that mostly we cannot statistically separate the groups (p-value > 0.05) for all values of *Thr* using only one of the calculated measures (except for $\langle L \rangle$, which is significant for $Thr > 0.65$, but its contribution to classification is low).

In the next step, we reduced the feature space by removing the features with a low contribution for the most threshold values—the average shortest path length $\langle L \rangle$ and the small-world coefficient σ. Thus, we used only three features for the LDA model training and classification: mean node strength $\langle k \rangle$, number of edges N_e, and clustering coefficient C. Figure 5(a) shows the classification accuracy in this case. Maximal accuracy is achieved for $Thr = 0$ (Accuracy $= 0.83 \pm 0.05$) and $Thr = 0.7$ (Accuracy $= 0.81 \pm 0.06$), while

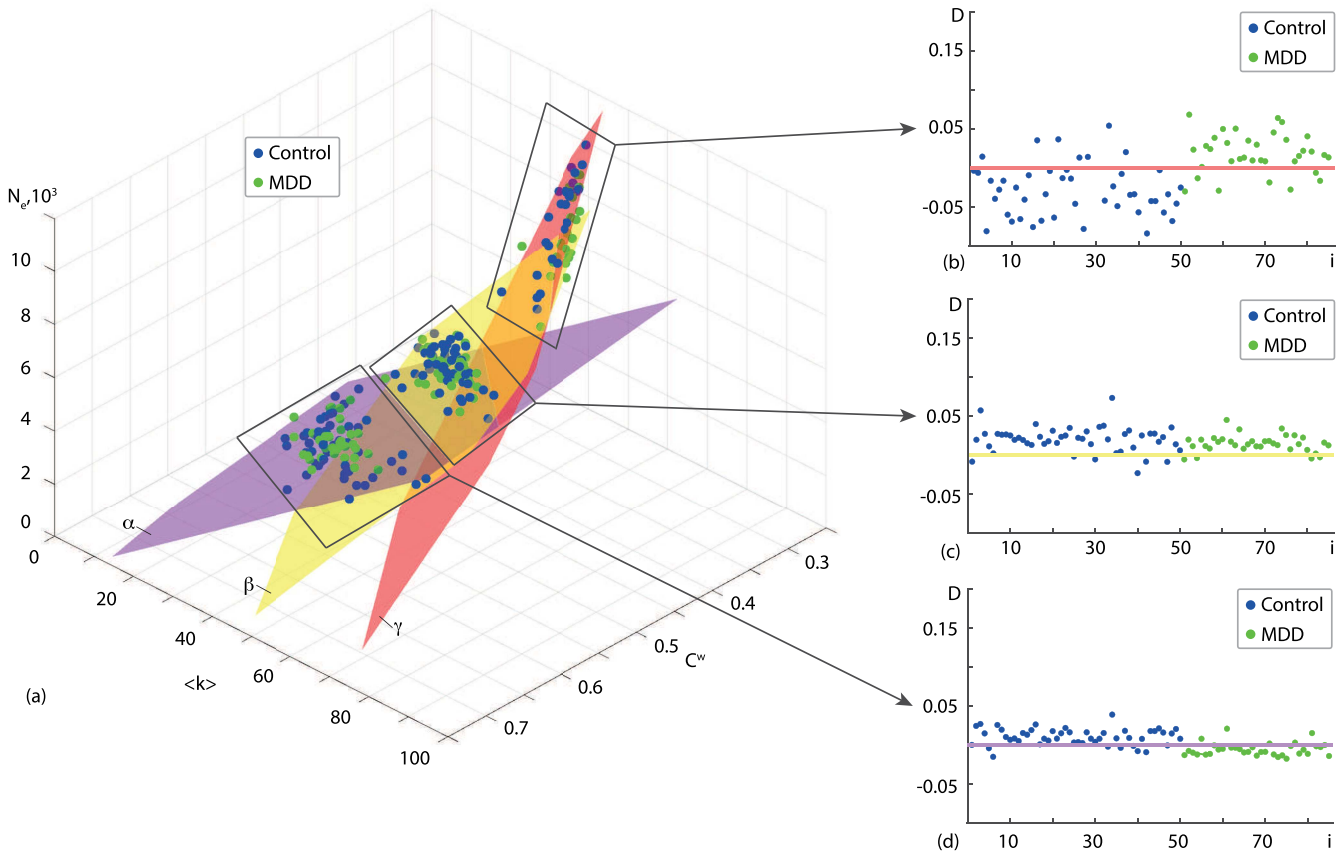

FIG. 6. (a) The network measures space for control (blue points) and MDD (green points) groups with the LDA separation planes (α, β, γ) and (b)–(d) the distances D between each ith subject (points) and the LDA separation plane for different threshold values: (b) $Thr = 0.65$ (γ plane), (c) $Thr = 0.45$ (β plane), and (d) $Thr = 0$ (α plane). Each point represents each subject.

minimal accuracy $= 0.56 \pm 0.07$ is achieved for $Thr = 0.45$. Comparing these results with Fig. 4(a), we can conclude that the accuracy of classification does not change sufficiently with a reduction in the number of features from 5 to 3. Then, we estimated the contribution of each feature to the classification outcome using the wrapper-type approach [see Fig. 5(b)]. For all threshold values, the contribution of two features correlates with accuracy: mean node strength $\langle k \rangle$ and clustering coefficient C for $Thr \in [0, 0.45]$ and mean node strength $\langle k \rangle$ and number of edges N_e for $Thr \in [0.45, 0.8]$. All of them have almost zero contribution for $Thr \in [0.4, 0.5]$ where the accuracy of classification is close to 0.5.

We used the MANOVA statistical test for the estimation of the statistical separability of control and MDD groups using the considered three features ($\langle k \rangle$, N_e, and C) together. The obtained results correlate with the capability of LDA to classify the groups [Fig. 5(c)]: for $0.36 < Thr < 0.53$, the p-value is higher than 0.05, and LDA cannot separate the control and MDD groups, while for other threshold values, p-value < 0.05 and accuracy of classification > 0.6.

Hence, using only one network measure, we cannot statistically separate the considered control and MDD groups, but when we introduce a set of measures, the groups become separable in the feature space. Using all couplings or only the strongest ones allows obtaining the maximal accuracy of classification, while using couplings with strength $w > 0.35$ leads to the inability of LDA to classify the experimental groups.

Finally, we analyze the network measures space (as feature space) depending on the threshold value [Fig. 6(a)]. Here, all subjects are depicted as points in the three-dimensional feature space (the control group is blue and MDD one is green) for three different threshold values $Thr = 0, 0.45, 0.65$. For each Thr, we plot the corresponding LDA separation plane (α, β, and γ, respectively) described by Eq. (6). For all participants, their measures lie approximately on the measures plane, and the planes for healthy and MDD groups are close to each other, but still, LDA is able to build a plane to separate them for the most number of cases. To illustrate the position of points relative to the plane for both groups, we calculate the distance from the point to the plane [Eq. (7)] for each subject for $Thr = 0.65$ [Fig. 6(b)], $Thr = 0.45$ [Fig. 6(c)], $Thr = 0$ [Fig. 6(d)]. As one can see, for the high threshold value [Fig. 6(b)], blue points (control subjects) are below the separation plane (purple line), and

green points (subjects with MDD) are above the plane, but for the low threshold, their relative to the plane locations are reversed [Fig. 6(d)]. For the middle value $Thr = 0.45$ [Fig. 6(c)], corresponding to the low accuracy of classification, both groups are shuffled, and it becomes impossible to separate them. More specifically, the parametric planes of two groups move and rotate in the feature space with the increasing threshold value [Fig. 6(a)], and their intersection increases with approaching $Thr = 0.45$, for which classification accuracy is minimal. It is caused by changing the network measures with the increasing threshold value. Due to all subjects lying on the measures plane regardless of the threshold value, we can suppose the connectivity between the measures.

IV. CONCLUSION

We have applied linear discriminant analysis for discrimination between patients with major depressive disorder and healthy controls by using functional connectivity networks derived from resting-state fMRI data. Unlike many previous papers,[34] we have utilized not the networks themselves but their global characteristics (measures): mean node strength, average shortest path length, number of edges, clustering coefficient, and small-world coefficient. We applied statistical methods in combination with the wrapper-type approach for feature selection and interpretation of the LDA model classification results.

These methods have shown that the classes are indistinguishable in the univariate feature space, i.e., individually, the network measures are close for the MDD and control groups. However, these classes become distinguishable in the three-dimensional feature space formed by the most important features (mean node strength, clustering coefficient, and the number of edges), which the proposed wrapper-type feature selection algorithm allowed to identify. The results of the statistical analysis confirmed the selection of these features as the most significant.

We have also investigated how the thresholding of connection weights in the functional networks (*Thr* parameter) influences classification accuracy. In the case of the three most significant features, we revealed that LDA achieves the highest accuracy (81%–83%) when we consider the network with all connections ($Thr = 0$) or only the strongest ones ($Thr > 0.7$). When the thresholding parameter lies in the range $[0.35, 0.55]$, the LDA model cannot separate the control and MDD groups, and the average accuracy is only 56%. Thus, we revealed the optimal ranges of the thresholding parameter.

In the three-dimensional space of the most significant measures, the points corresponding to individual subjects lie on the parametric plane, with the planes for control and MDD groups being close to each other. This explains the difficulty of classifying the considered classes. Moreover, this result means that a linear combination of the most significant features allows us to distinguish the considered classes. We revealed that the parametric planes of two groups rotate in the feature space with increasing the thresholding parameter, and their intersection increases with approaching $Thr = 0.45$, for which classification accuracy is minimal.

Thus, the proposed combined approach solved the problem of selecting the most significant features and the optimal thresholding parameter to achieve high accuracy of MDD patients' discrimination while allowing full interpretability of the results and the classifier operation. Moreover, this approach allowed us to analyze the separability of classes in the multidimensional feature space.

The proposed approach can be applied to other ML tasks when classifying the complex networks. It allows us to estimate the most important network features for the maximal separability of the groups depending on the external parameter. The results of the application of the proposed approach are easy to interpret because a user always knows which features lead to achieving high accuracy, and their further analysis allows understanding the nature of the compared groups. Particularly, one can estimate the number of the most different features between healthy and unhealthy brain networks in our case.

ACKNOWLEDGMENTS

The work was funded by the Russian Science Foundation (Grant No. 23-71-30010).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Andrey V. Andreev: Conceptualization (equal); Investigation (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Semen A. Kurkin:** Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Validation (equal); Writing – review & editing (equal). **Drozdstoy Stoyanov:** Data curation (equal); Investigation (equal); Writing – review & editing (equal). **Artem A. Badarin:** Formal analysis (equal); Software (equal); Validation (equal). **Rossitsa Paunova:** Data curation (equal); Investigation (equal); Resources (equal); Validation (equal). **Alexander E. Hramov:** Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]J. Wang, X. Zuo, and Y. He, "Graph-based network analysis of resting-state functional MRI," Front. Syst. Neurosci. **4**, 16 (2010).
[2]E. Bullmore and O. Sporns, "The economy of brain network organization," Nat. Rev. Neurosci. **13**(5), 336–349 (2012).
[3]H.-J. Park and K. Friston, "Structural and functional brain networks: From connections to cognition," Science **342**(6158), 1238411 (2013).
[4]R. Wang, M. Liu, X. Cheng, Y. Wu, A. Hildebrandt, and C. Zhou, "Segregation, integration, and balance of large-scale resting brain networks configure different cognitive abilities," Proc. Natl. Acad. Sci. U.S.A. **118**(23), e2022288118 (2021).
[5]A. E. Hramov, N. S. Frolov, V. A. Maksimenko, S. A. Kurkin, V. B. Kazantsev, and A. N. Pisarchik, "Functional networks of the brain: From connectivity restoration to dynamic integration," Phys.-Usp. **64**(6), 584 (2021).

[6]M. Perovnik, T. Rus, K. A. Schindlbeck, and D. Eidelberg, "Functional brain networks in the evaluation of patients with neurodegenerative disorders," Nat. Rev. Neurol. **19**(2), 73–90 (2023).
[7]S. Kurkin, N. Smirnov, E. Pitsik, M. S. Kabir, O. Martynova, O. Sysoeva, G. Portnova, and A. Hramov, "Features of the resting-state functional brain network of children with autism spectrum disorder: EEG source-level analysis," Eur. Phys. J. Spec. Top. **232**, 683–693 (2023).
[8]T. Yamada, R.-I. Hashimoto, N. Yahata, N. Ichikawa, Y. Yoshihara, Y. Okamoto, N. Kato, H. Takahashi, and M. Kawato, "Resting-state functional connectivity-based biomarkers and functional MRI-based neurofeedback for psychiatric disorders: A challenge for developing theranostic biomarkers," Int. J. Neuropsychopharmacol. **20**(10), 769–781 (2017).
[9]F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608 (2017).
[10]Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," Multimed. Syst. **28**(6), 2335–2355 (2022).
[11]Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," Diagnostics **12**(2), 237 (2022).
[12]O. E. Karpov, V. V. Grubov, V. A. Maksimenko, S. A. Kurkin, N. M. Smirnov, N. P. Utyashev, D. A. Andrikov, N. N. Shusharina, and A. E. Hramov, "Extreme value theory inspires explainable machine learning approach for seizure detection," Sci. Rep. **12**(1), 11474 (2022).
[13]S. Kundu, "Ai in medicine must be explainable," Nat. Med. **27**(8), 1328 (2021).
[14]A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," Neural Comput. Appl. **32**(24), 18069–18083 (2020).
[15]O. E. Karpov, E. N. Pitsik, S. A. Kurkin, V. A. Maksimenko, A. V. Gusev, N. N. Shusharina, and A. E. Hramov, "Analysis of publication activity and research trends in the field of ai medical applications: Network approach," Int. J. Environ. Res. Public Health **20**(7), 5335 (2023).
[16]A. Jha, J. K. Aicher, M. R. Gazzara, D. Singh, and Y. Barash, "Enhanced integrated gradients: Improving interpretability of deep learning models using splicing codes as a case study," Genome Biol. **21**(1), 149 (2020).
[17]H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, and T. E. Arnoff, "Biologically informed deep neural network for prostate cancer discovery," Nature **598**(7880), 348–352 (2021).
[18]S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, and M. Kaku, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," Brain **143**(6), 1920–1933 (2020).
[19]C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell. **1**(5), 206–215 (2019).
[20]O. E. Karpov, M. S. Khoymov, V. A. Maksimenko, V. V. Grubov, N. Utyashev, D. A. Andrikov, S. A. Kurkin, and A. E. Hramov, "Evaluation of unsupervised anomaly detection techniques in labelling epileptic seizures on human EEG," Appl. Sci. **13**(9), 5655 (2023).
[21]C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—A brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, 14–18 September 2020, Proceedings* (Springer, 2021), pp. 417–431.
[22]A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **9**(4), e1312 (2019), see https://pubmed.ncbi.nlm.nih.gov/32089788/.
[23]S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability," Soft Comput. **13**, 959–977 (2009).
[24]I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications* (Springer, 2008), Vol. 207.
[25]Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics **23**(19), 2507–2517 (2007).
[26]M. Zanin, P. Sousa, D. Papo, R. Bajo, J. García-Prieto, F. D. Pozo, E. Menasalvas, and S. Boccaletti, "Optimizing functional network representation of multivariate time series," Sci. Rep. **2**(1), 630 (2012).
[27]A. Kuc, S. Korchagin, V. A. Maksimenko, N. Shusharina, and A. E. Hramov, "Combining statistical analysis and machine learning for EEG scalp topograms classification," Front. Syst. Neurosci. **15**, 716897 (2021).
[28]N. Frolov, M. S. Kabir, V. Maksimenko, and A. Hramov, "Machine learning evaluates changes in functional connectivity under a prolonged cognitive load," Chaos **31**(10), 101106 (2021).
[29]P. Zachar, D. S. Stoyanov, M. Aragona, and A. Jablensky, *Alternative Perspectives on Psychiatric Validation* (Oxford University Press, Oxford, 2014).
[30]A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, and A. F. Schatzberg, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," Nat. Med. **23**(1), 28–38 (2017).
[31]D. S. Stoyanov, R.-D. Stieglitz, C. Lenz, and S. Borgwardt, "The translational validation as novel approach to integration of neuroscience and psychiatry," in *New Developments in Clinical Psychology Research* (Nova Science, 2015), pp. 196–208.
[32]A. Todeva-Radneva, R. Paunova, S. Kandilarova, and D. St. Stoyanov, "The value of neuroimaging techniques in the translation and transdiagnostic validation of psychiatric diagnoses-selective review," Curr. Top. Med. Chem. **20**(7), 540–553 (2020).
[33]L.-L. Zeng, H. Shen, L. Liu, and D. Hu, "Unsupervised classification of major depression using functional connectivity MRI," Hum. Brain Mapp. **35**(4), 1630–1641 (2014).
[34]D. Stoyanov, V. Khorev, R. Paunova, S. Kandilarova, D. Simeonova, A. Badarin, A. Hramov, and S. Kurkin, "Resting-state functional connectivity impairment in patients with major depressive episode," Int. J. Environ. Res. Public Health **19**(21), 14045 (2022).
[35]Y. Li, X. Dai, H. Wu, and L. Wang, "Establishment of effective biomarkers for depression diagnosis with fusion of multiple resting-state connectivity measures," Front. Neurosci. **15**, 729958 (2021).
[36]E. N. Pitsik, V. A. Maximenko, S. A. Kurkin, A. P. Sergeev, D. Stoyanov, R. Paunova, S. Kandilarova, D. Simeonova, and A. E. Hramov, "The topology of fMRI-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder," Chaos Soliton. Fract. **167**, 113041 (2023).
[37]M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," NeuroImage **145**, 137–165 (2017).
[38]A. Lord, D. Horn, M. Breakspear, and M. Walter, "Changes in community structure of resting state functional connectivity in unipolar depression," PLoS One **7**(8), e41282 (2012).
[39]E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," NeuroImage **206**, 116189 (2020).
[40]A. M. Bastos and J.-M. Schoffelen, "A tutorial review of functional connectivity analysis methods and their interpretational pitfalls," Front. Syst. Neurosci. **9**, 175 (2016).
[41]M. Rubinov and O. Sporns, "Weight-conserving characterization of complex functional brain networks," NeuroImage **56**(4), 2068–2079 (2011).
[42]M. P. Van Den Heuvel, C. J. Stam, R. S. Kahn, and H. E. H. Pol, "Efficiency of functional brain networks and intellectual performance," J. Neurosci. **29**(23), 7619–7624 (2009).
[43]D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**(6684), 440–442 (1998).
[44]G. Costantini and M. Perugini, "Generalization of clustering coefficients to signed correlation networks," PLoS One **9**(2), e88669 (2014).
[45]M. D. Humphries and K. Gurney, "Network 'small-world-ness': A quantitative method for determining canonical network equivalence," PLoS One **3**(4), e0002051 (2008).
[46]R. O. Duda and P. E. Hart, *Pattern Classification* (John Wiley & Sons, 2006).