frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 21 July 2025
DOI 10.3389/fnins.2025.1609547

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Ju Gao,
Suzhou Guangji Hospital, China
Aviral Chharia,
Carnegie Mellon University, United States

*CORRESPONDENCE
Shaolong Wei
✉ weishaolong37@gmail.com
Hongcheng Yao
✉ yaohongcheng19@gmail.com

RECEIVED 10 April 2025
ACCEPTED 01 July 2025
PUBLISHED 21 July 2025

CITATION
Yuan X, Wei S, Sun Y, Gu L, He Y, Chen T, Yao H and Rao H (2025) Robust multi-task feature selection with counterfactual explanation for schizophrenia identification using functional brain networks. *Front. Neurosci.* 19:1609547. doi: 10.3389/fnins.2025.1609547

Robust multi-task feature selection with counterfactual explanation for schizophrenia identification using functional brain networks

Xinyan Yuan[1], Shaolong Wei[2]*, Ying Sun[1], Lingling Gu[1], Yanyan He[1], Tiantian Chen[1], Hongcheng Yao[3]* and Haonan Rao[3]

[1]School of Electronics and Information, Jiangsu Vocational College of Business, Nantong, China, [2]School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China, [3]School of Information Science and Technology, Nantong University, Nantong, China

Introduction: Functional brain networks measured by resting-state functional magnetic resonance imaging (rs-fMRI) have become a promising tool for understanding the neural mechanisms underlying schizophrenia (SZ). However, the high dimensionality of these networks and small sample sizes pose significant challenges for effective classification and model generalization.

Methods: We propose a robust multi-task feature selection method combined with counterfactual explanations to improve the accuracy and interpretability of SZ identification. rs-fMRI data are preprocessed to construct a functional connectivity matrix, and features are extracted by sorting the upper triangular elements. A multi-task feature selection framework based on the Gray Wolf Optimizer (GWO) is developed to identify abnormal functional connectivity (FC) features in SZ patients. A counterfactual explanation model is applied to reduce perturbations in abnormal FC features, returning the model prediction to normal and enhancing clinical interpretability.

Results: Our method was tested on five real-world SZ datasets. The results demonstrate that the proposed method significantly outperforms existing methods in terms of classification accuracy while offering new insights into the analysis of SZ through improved feature selection and explanation.

Discussion: The integration of multi-task feature selection and counterfactual explanation improves both the accuracy and interpretability of SZ identification. This approach provides valuable clinical insights by revealing the key functional connectivity features associated with SZ, which could assist in the development of more effective diagnostic tools.

KEYWORDS

schizophrenia, functional connectivity, rs-fMRI, feature selection, counterfactual explanation

1 Introduction

Schizophrenia (SZ) is a chronic, often disabling mental disorder that affects one percent of the world's population (Insel, 2010; McCutcheon et al., 2020). Patients' clinical symptoms manifest in perception, thinking, and emotion, such as hallucinations, delusions, incoordinated excitement, and anxiety (Song et al., 2023; Rantala et al., 2022). Although the pathogenesis of SZ is still unclear, it is increasingly recognized that analyzing the brain network of SZ can help improve differential diagnosis and understand the pathological mechanism (Zhang et al., 2021). Recent studies have shown that functional

brain networks measured by resting-state functional magnetic resonance imaging (rs-fMRI) have become a promising tool to reveal the underlying neural mechanisms of SZ (Zhu et al., 2024; Chyzhyk et al., 2015). SZ causes widespread changes in functional brain networks, including changes in global brain topology, abnormal connectivity in local regions, and the formation of specific abnormal subgraphs (Huang et al., 2025).

However, although functional brain networks provide rich pathological information, these data often have high-dimensional characteristics, making analysis and modeling face great challenges (Mhiri and Rekik, 2020). Therefore, feature selection (FS) becomes an indispensable step, which can remove irrelevant or redundant features and retain only the most diagnostically valuable information (Naheed et al., 2020). In addition, functional brain network data usually face the problem of small samples. Due to the high cost of data acquisition, the long experimental cycle, and the difficulty in recruiting subjects, the number of samples is often much lower than the feature dimension, making model training susceptible to overfitting, thereby reducing generalization ability (Turner et al., 2018; Ding et al., 2024). In this context, robust and effective FS is vital. In fact, FS plays a key role in identifying meaningful biomarkers, such as functional connectivity between brain regions, which can characterize abnormalities in brain function associated with brain diseases such as SZ, thus providing insight into understanding the neural basis of brain diseases, as well as diagnosis and prediction (Xing et al., 2022).

For functional brain network data, the traditional FS method often exhibits poor robustness across datasets, primarily due to the high dimensionality of the feature space and the scarcity of training samples, and it is difficult to identify connection features with consistency and biological interpretability (Wang et al., 2015; Lv et al., 2015; Hu et al., 2021). At present, most existing FS methods have combined advanced technologies such as machine learning or deep learning to improve performance, such as using graph neural networks to model FC structures, or improving feature selection efficiency through embedded FS strategies, but these methods still have obvious limitations. On the one hand, many models still lack consistent evaluation across data sets, making it difficult to identify robust disease-related connection features (Chan et al., 2024); on the other hand, most existing methods are black-box in form and lack interpretability, especially in clinical applications. It is difficult to provide actionable explanations or intervention recommendations (Verma et al., 2023). In addition, although some studies have introduced multimodal or high-order connection features in SZ diagnosis, it is still difficult to achieve a good balance between model generalization and explanatory power (Sunil et al., 2024).

To address the above challenges and fill this gap, we proposed a novel and robust multi-task feature selection method for SZ diagnosis, and explained the changes in brain functional connectivity (FC) caused by the disease through a counterfactual explanation model. The schematic diagram of our proposed method is shown in Figure 1. Specifically, we first preprocessed the rs-fMRI data, constructed the FC matrix, and then extracted the upper triangular elements as feature vectors and sorted them. Subsequently, we developed a robust multi-task feature selection framework based on the Gray Wolf Optimizer (GWO), and selected the abnormal FC features of SZ patients by adopting feature stratification and weight-based task generation. Finally, we used the counterfactual explanation model to generate a set of counterfactual examples for SZ patients, that is, by fine-tuning the abnormal FC features of SZ patients to make their state close to normal, thus providing theoretical guidance for the analysis and diagnosis of SZ. We verified the effectiveness of our method on five real SZ datasets, and the results showed that our method not only improved the interpretability of the model, but also provided a new perspective for the analysis of SZ. The main contributions of this paper are as follows:

- We propose a Robust Multi-Task Feature Selection with Counterfactual Explanation for Schizophrenia Identification to assist SZ analysis and diagnosis.
- We construct a multi-task feature selection framework based on GWO and combine it with the counterfactual explanation model to fine-tune the abnormal FC features of SZ patients to make their status closer to that of healthy individuals, thereby improving the accuracy of SZ classification and the interpretability of the model.
- We evaluate the performance of the proposed method using five real SZ datasets. The results show that the proposed method outperforms existing methods.

2 Related work

2.1 Gray wolf optimizer

Gray Wolf Optimizer (GWO) (Mirjalili et al., 2014) is an intelligent optimization algorithm that simulates the hunting behavior of gray wolf groups. In the context of multitasking, GWO provides efficient global search capabilities and information-sharing mechanisms between individuals, which can improve optimization performance in a multi-task environment.

Gray wolf packs are generally divided into four levels: (i) α is the leader of the wolf pack, representing the current optimal solution, (ii) β is the second-level wolf, assisting α in decision-making, representing the second-best solution, (iii) δ is the third-level wolf, assisting β, representing the third-best solution, and (iv) θ is an ordinary wolf that obeys other high-level wolves and represents the remaining candidate solutions. When searching for prey, gray wolves will gradually approach the prey and surround it:

$$D = |C \cdot X_p - X| \tag{1}$$

$$X(t+1) = X_p - A \cdot D \tag{2}$$

where X_p is the location of the prey or the current optimal solution, X is the location of the individual wolf, t is the number of iterations, and A and C are coefficient vectors, which are calculated as follows:

$$A = 2d \cdot r_1 - d, \; C = 2r_2 \tag{3}$$

where d is the convergence factor that decreases linearly with the number of iterations, from 2 to 0, and r_1 and r_2 are random numbers between $[0, 1]$. GWO uses three optimal solutions (α, β, δ) to jointly guide the search:

$$X(t+1) = \frac{1}{3} \sum_{i=\alpha,\beta,\delta} (X_i - A_i \cdot D_i) \tag{4}$$

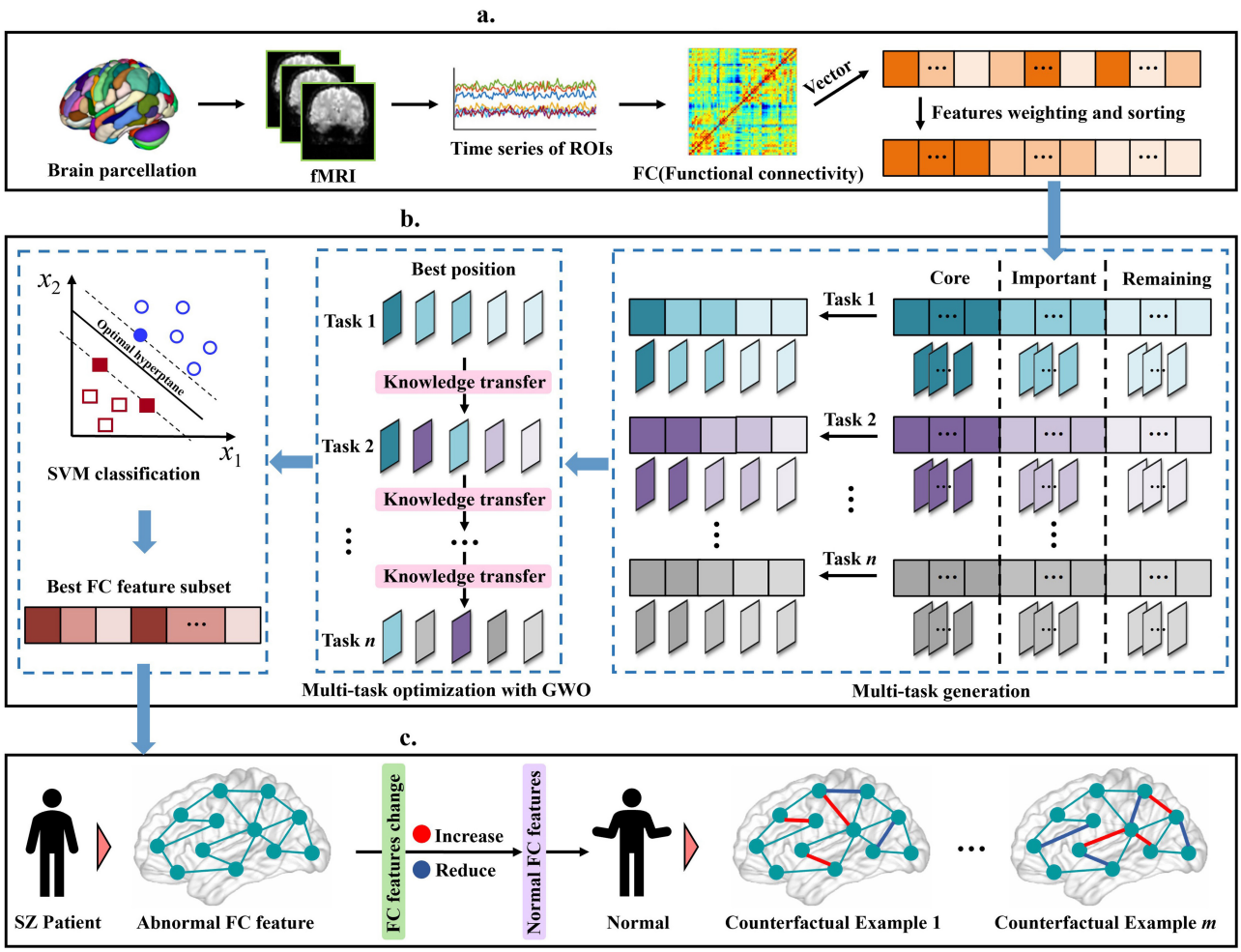

FIGURE 1
Illustration of our proposed schizophrenia analysis method, including **(a)** data pre-processing, **(b)** Robust Multi-Task Feature Selection, **(c)** Diversity counterfactual explanation.

where $D_i = |C_i \cdot X_i - X|, i \in \{\alpha, \beta, \delta\}$. When $|A|$ becomes smaller (approaches 0), the search range is reduced, and the wolf pack gradually converges to the optimal solution. When $|A| > 1$, the wolf pack stays away from the prey and performs a global search to avoid falling into the local optimum.

2.2 Counterfactual explanation

Counterfactual explanations are a method for making machine learning models more transparent by showing how to change attributes to obtain different results (Spreitzer et al., 2022). Cheng et al. (2020) introduced counterfactuals with a classic example: A person submitted a loan request but was rejected by the bank. If his credit score had been 700 instead of 600, his loan application would have been approved.

Counterfactual explanations are currently widely used in different fields, including medical diagnosis, decision reasoning, and artificial intelligence. Richens et al. (2020) have improved the application of machine learning in the field of medical diagnosis, especially in identifying rare diseases, by establishing a counterfactual causal diagnosis model. Prado-Romero et al. (2023) use counterfactual explanations to provide a way to understand model decisions by providing specific changes in input features to explain the model's decision-making process. In addition, counterfactual explanations also have many applications in brain networks. For example, in the study of Abrate and Bonchi (2021), they proposed an explanation method for a black-box graph classifier for brain network classification. By analyzing counterfactual graphs, brain region connection patterns associated with specific brain region diseases can be identified. Matsui et al. (2022) proposed a new generative deep neural network (DNN) called Counterfactual Activation Generator to provide counterfactual explanations for DNN-based brain activation classifiers.

Counterfactual explanation has emerged as an important branch in the field of machine learning interpretability; however, it has not yet been applied to FC analysis. In this work, we introduce a counterfactual perspective: if the abnormal FC between brain regions in SZ patients is adjusted toward the normal range, their

predicted state may shift closer to that of healthy individuals. Such counterfactual reasoning is particularly valuable in the medical domain, as it can assist clinicians in evaluating the potential impact of different treatment strategies, especially in the context of brain diseases.

3 Materials and methods

3.1 Schizophrenia dataset

In this study, five public datasets are used, including the Center for Biomedical Research (COBRE) dataset (120 subjects), the Huaxi dataset (311 subjects), the Nottingham dataset (68 subjects), the Taiwan dataset (131 subjects) and the Xiangya dataset (143 subjects). All subjects met the following conditions: (i) no other Diagnostic and Statistical Manual of Mental Disorders (DSMIV) disease exists, (ii) no history of drug abuse, (iii) no clinically significant head trauma. The specific information of the subjects is presented in Table 1.

3.2 Data pre-processing

The rs-fMRI data of the five datasets are collected by different types of scanners, including COBRE and Xiangya by 3-T Siemens Tim-Trio scanner with an eight or 12-channel head coil, Huaxi by 3-T General Electric MRI scanner, and Nottingham by 3-T Philips Achieva MRI scanner. The rs-fMRI data are preprocessed using the program standard procedures of SPM 8 and the Data Processing Assistant for Resting-State fMRI (DPARSF). The following steps are performed: (i) removing the first 10 volumes, (ii) slice timing correction, (iii) head motion correction, (iv) regress out the nuisance covariates, (v) normalized to standardized space, (vi) voxel-wise bandpass filtering, (vii) normalization of anatomical images to MNI template space, and (viii) smoothing with a 4 mm Full Width at Half Maximum (FWHM) Gaussian kernel. After processing, we defined the nodes of the brain network according to the Automatic Anatomical Labeling (AAL) template, and calculated the pairwise similarities between the noded1s of the time series as the connecting edges of the brain network.

Next, let $A_i^F \in \mathbb{R}^{N\times N}$ be the connectivity matrix of the functional brain network, N be the number of regions of the brain network, $i = 1, 2, ..., p$, and p be the number of subjects. We take the upper triangular elements of the matrix as features and represent them as vectors $S_i = (s_i^1, ..., s_i^j, ..., s_i^q) \in \mathbb{R}^{1\times q}$, $q = \frac{N(N-1)}{2}$, s_i^j represents the j-th feature of the i-th subject, and $Y_i \in \mathbb{R}$ is the label of the i-th subject. It is worth noting that in this paper, we divided the brain network into 90 regions of interest (ROI), that is, $N = 90$, so each subject contains a vector of dimension $1 \times 4{,}005$, which reflects the functional connectivity strength pattern between the 90 brain regions of the subject.

3.3 Robust multi-task feature selection

3.3.1 Multi-task generation

To identify the most critical FC features for brain disease diagnosis, we use the infinite feature selection (IFS) (Roffo et al., 2020) method to calculate the importance of each feature and rank the features accordingly. Specifically, the weight of each feature is calculated based on the linear weighting of the following three aspects (i.e., Fisher criterion h_j, mutual information m_j, and standard deviation σ_j). The first is the Fisher criterion:

$$h_j = \frac{|\mu_{j,1} - \mu_{j,2}|^2}{\sigma_{j,1}^2 + \sigma_{j,2}^2} \tag{5}$$

where $\mu_{j,g}$ and $\sigma_{j,g}$ represent the mean and standard deviation of the j-th feature in the g-th class, respectively. In our experiments, both are binary classifications, so $g \in \{0, 1\}$.

The second is the normalized mutual information m_j between feature s^j and class label Y:

$$m_j = \sum_{y\in Y}\sum_{z\in s^j} u(z, y) log(\frac{u(z, y)}{u(z)u(y)}) \tag{6}$$

where Y is the set of class labels and $u(\cdot)$ represents the joint distribution probability.

TABLE 1 Characteristics of subjects in the five datasets in this study.

Datasets	Class	Gender (M/F)	*P*-value of gender	Age (years)	*P*-value of age
COBRE	NC	46/21	0.1927	34.82+11.28	0.3987
	SZ	42/11		36.75+13.68	
Huaxi	NC	79/71	0.6748	27.80+12.50	1.000
	SZ	80/81		27.80+12.50	
Nottingham	NC	26/10	0.2277	33.38+8.98	0.9855
	SZ	27/5		33.34+9.05	
Taiwan	NC	25/37	0.2329	29.87+8.62	0.2847
	SZ	35/34		31.59+9.60	
Xiangya	NC	35/25	0.9333	27.17+6.64	0.1025
	SZ	49/34		23.37+7.83	

NC, normal control; SZ, schizophrenia.

The third is the standard deviation σ_j, which reflects the dispersion of feature s^j in the sample.

The final weight of each feature s_i is calculated as follows:

$$s_i = \alpha_1 \cdot h_j + \alpha_2 \cdot m_j + \alpha_3 \cdot \sigma_j. \tag{7}$$

where $\alpha_1+\alpha_2+\alpha_3 = 1$, this weighting approach allows us to flexibly adjust the contribution of each indicator in the selection of features, thus selecting the most informative features for the diagnosis of schizophrenia (SZ).

Based on the preliminary evaluation of FC feature importance based on the above three factors, we further constructed a feature weight curve and optimized the FS process by introducing a knee point detection algorithm, following the knee point detection method proposed by Chen et al. (2021). This approach provides an automated criterion for determining the optimal feature subset size. Specifically, after obtaining the weight of each feature, we first construct a straight line connecting the starting point and the end point of the weight curve, and then calculate the vertical distance from each point on the curve to the straight line. The knee point (x_{knee}, y_{knee}) is the point that maximizes the distance:

$$(x_{knee}, y_{knee}) = \arg\max_j \left(\frac{|y_j - (ax_j + b)|}{\sqrt{a^2 + 1}}\right) \tag{8}$$

where a and b are the slope and intercept of the straight line determined by the starting point and the end point, (x_j, y_j) is the coordinate of the j-th feature point on the curve, $j = 2, 3, ..., q-1$. The identified knee points divide the feature weight curve into multiple intervals, and the features in each interval are given different priorities according to their weights.

Based on the location of the knee points, as shown in Figure 1b, we divide the features into three categories:

(i) Core features: located before the first knee point. These features are usually highly correlated with the predicted target variable and have low redundancy, and contribute the most to the model's predictive ability.

(ii) Important features: located between the two knee points. Although these features are not as important as the core features, may still contain useful information for specific scenarios. When combined with other features, they can enhance overall model performance, especially in complex cases where feature interactions are significant.

(iii) Remaining features: located after the second knee point. These features contribute less to the prediction task, contain redundant information, or have low correlation with the target variable.

After the above steps, we further use this category information to guide the task generation process. To ensure that the feature extraction process not only reflects its relative importance but also maintains appropriate diversity, we adopt a probabilistic extraction method based on feature weights. Specifically, we determine the initial selection probability of each feature based on the feature weight.

$$P_j = \frac{\omega_j}{\sum_{j=1}^{q} \omega_j} \tag{9}$$

where ω_j is the weight of the j-th feature. The larger ω_j is, the higher its initial extraction probability is, and thus it is given priority in FS. To ensure that all features have a certain chance of being selected and to avoid the extraction probability of low-weight features becoming too small, we adjust the initial probability:

$$P'_j = \frac{P_j}{\max(P_j)} \tag{10}$$

The above formula ensures that the maximum extraction probability of a feature is 1, and the extraction probabilities of all other features are adjusted proportionally, avoiding excessive neglect of low-weight features while still maintaining the priority of high-weight features during extraction.

During the task generation process, a random number λ between 0 and 1 is first randomly generated, which is used to determine which features will be selected for the current task. For each feature s^j, if $\lambda \leq P'_j$, the feature will be selected for the current task. As shown in Figure 1b, after n rounds of independent extraction, n different task sets are generated, each of which contains a set of selected feature subsets. This mechanism ensures that high-weight features are selected first and fully retain the potential contribution of low-weight features, thereby effectively improving the diversity and flexibility of the task generation process.

3.3.2 Multi-task optimization with GWO

In multi-task optimization, we propose to combine the knowledge transfer mechanism with the GWO-based multi-task optimization method to enhance information sharing between different tasks, thereby improving the efficiency and effect of overall optimization. Specifically, we directly integrate the knowledge transfer mechanism in the initialization phase of GWO to make full use of the optimization experience of existing tasks.

To achieve effective knowledge transfer, in the multi-task optimization process, we first need to quantify the importance of each feature in the previous task. In other words, we need to calculate the cumulative number of times Q_{KT} that feature s^j is selected in all previous tasks:

$$Q_{KT}(s^j) = \sum_{t=1}^{n} Q_{KT}^t(s^j) \tag{11}$$

where n represents the total number of tasks, $Q_{KT}^t(s^j)$ represents whether the feature is selected in the t-th task (if selected, it is 1, otherwise it is 0). Then, calculate the probability $P(s^j)$ of feature s^j being selected in the initial population of the new task:

$$P(s^j) = \frac{Q_{KT}(s^j)}{\sum_{j=1}^{q} Q_{KT}(s^j)} \tag{12}$$

The above formula converts the historical performance of the feature into a probability value, which will be directly applied to initialize the wolf pack:

$$G_{wo} = \begin{cases} 1, & \lambda \leq P(s^j) \\ 0, & \lambda > P(s^j) \end{cases} \tag{13}$$

where the random number $\lambda \in [0, 1]$, the feature s^j is selected only when it is less than or equal to $P(s^j)$. For ease

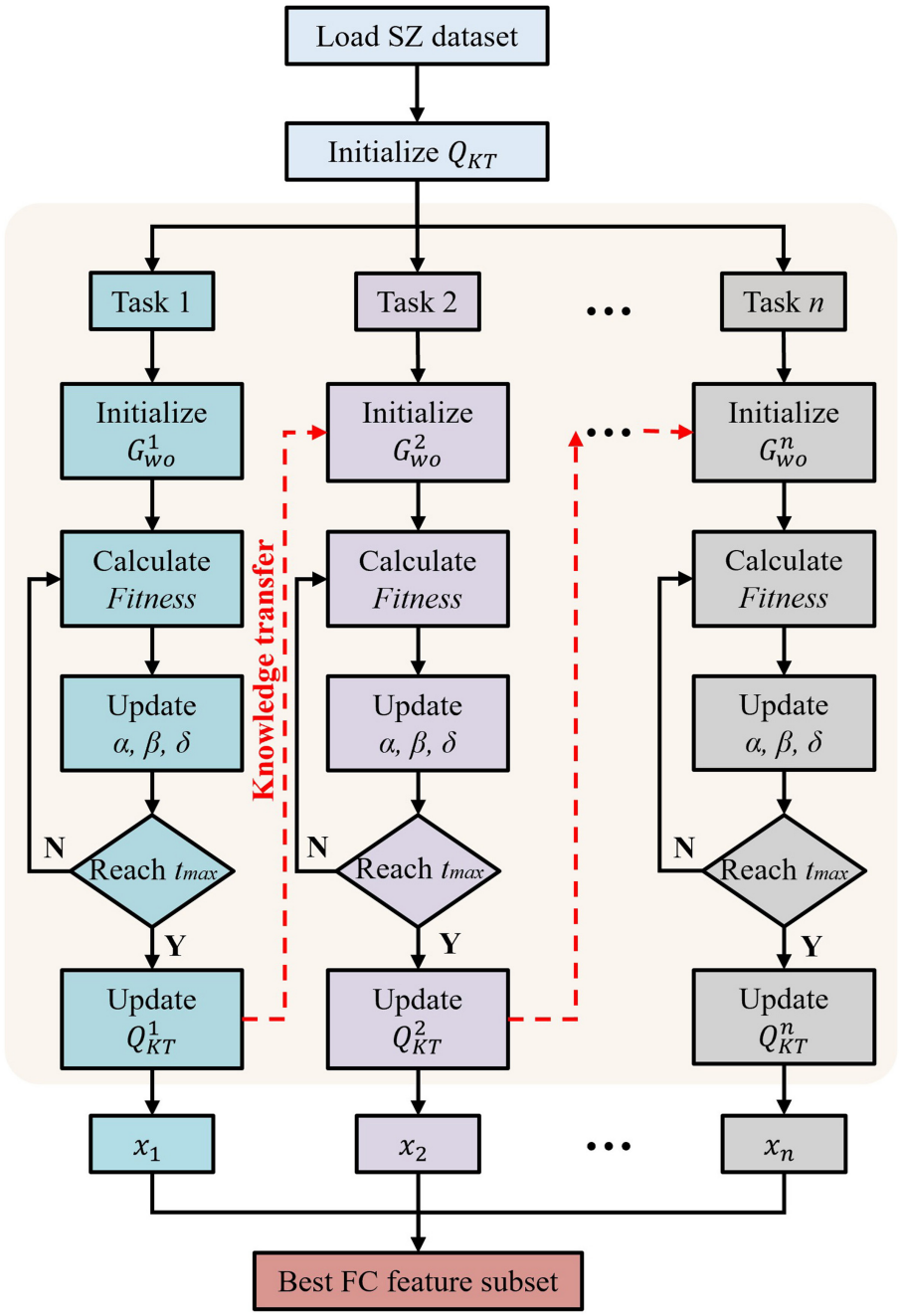

FIGURE 2
Flowchart of proposed multi-task optimization with GWO.

of understanding, we show the specific process of the proposed multi-task optimization method in Figure 2. First, the global environment is set. Subsequently, the algorithm enters a loop and processes n tasks in turn. For each task, the wolf pack is initialized independently, using the global knowledge of the previously processed tasks to provide information for the initial state of the search for the new task. The position of the wolf is iteratively updated to optimize the FS problem. After optimization, the best solution is used to update the global knowledge base. This cycle is repeated for each task, ensuring the continuous flow of information and the improvement of the solution. Finally, n feature subsets $(x_1, x_2, ..., x_n)$ are obtained from the n tasks.

In addition, to minimize the number of selected features while maintaining a high classification accuracy, we designed a fitness function in multi-task optimization and introduced a penalty term to constrain the number of features:

$$Fitness = \rho \times ACC - (1 - \rho) \times \frac{q_{sf}}{q} \tag{14}$$

where ρ is a weight coefficient, which ranges between $[0, 1]$ and is used to balance the classification accuracy ACC and the number of selected features q_{sf}.

After the above operations, we represent the selected feature matrix as $S' \in \mathbb{R}^{p\times k}$, where $k \ll q$. Based on the selected feature matrix S', we can train a suitable machine learning model [i.e., $f(\cdot)$] to predict schizophrenia. In our experiment, since the support vector machine (SVM) is strongly adaptable to small sample data sets, we used SVM as the classification model.

3.3.3 Diversity counterfactual explanation

To enhance the interpretability of our method, we further introduce a counterfactual explanation model (Mothilal et al., 2020) to generate sample-level explanations. The input of this model includes a trained SVM model [i.e., $f(\cdot)$] and the feature vector $c_i \in \mathbb{R}^{1\times k}$ of the i-th subject. Our goal is to generate a set of counterfactual examples $\{x_i^1, x_i^2, ..., x_i^L\}$ for subject i such that its decision outcome $x_i^l \in \mathbb{R}^{1\times k}$ is different from the prediction of the original feature vector c_i.

The counterfactual explanation model consists of three parts: loss function $loss(\cdot)$, distance function $dist(\cdot)$, and diversity metric $diversity(\cdot)$. Specifically, the first part pushes counterfactual x_i^l toward different predictions, the second part makes counterfactual examples closer to the original input, and the third part is used to increase the diversity of counterfactual explanations. In the first part, we use a hinge loss function that helps generate counterfactuals with less variation by reducing the preference for extreme values. The hinge loss is expressed as follows:

$$loss_{hinge} = max(0, 1 - z \cdot logit(f(x)) \tag{15}$$

where z is 1 when $\hat{Y} = 1$ and -1 when $\hat{Y} = 0$, and $logit(f(x))$ is the unscaled output of the SVM model. It is worth noting that in our experiments, 1 corresponds to normal subjects and 0 corresponds to patients, so in the verification of converting patients into normal subjects, $\hat{Y}$ is usually set to 1. For the choice of distance function in the second part, we follow Wachter et al. (2017) proposal and divide the distance of each feature by the median absolute deviation (MAD) of the feature values in the training set:

$$dist(x, c) = \frac{1}{L}\sum_{\alpha=1}^{L}\frac{|x^\alpha - c^\alpha|}{MAD_\alpha} \tag{16}$$

where MAD_α is the median absolute deviation of the α-th feature, L is the total number of counterfactual examples to generate, x represents the counterfactual example and c represents the original feature vector. For the third part, we use a determinant-based point procedure to measure the diversity of counterfactual examples, computed by the determinant value of its kernel matrix K:

$$diversity = det(K) \tag{17}$$

where $K_{u,v} = \frac{1}{1+dist(x^u, x^v)}$, x^v and x^u represent two counterfactual examples. In the experiments, to avoid uncertain determinants, we add small random perturbations on the diagonal elements to calculate the determinant.

Finally, we can obtain counterfactual examples by optimizing the following loss:

$$\begin{aligned} X(c_i) = {} & \frac{\gamma_1}{L}\sum_{l=1}^{L} dist(x_i^l, c_i) \\ & -\gamma_2 diversity(x_i^1, x_i^2, ..., x_i^L) \\ & +\arg\min_{x_i^1, x_i^2, ..., x_i^L} \frac{1}{L}\sum_{l=1}^{L} loss_{hinge}(f(x_i^l), \hat{Y}) \end{aligned} \tag{18}$$

where $X(c_i)$ is the final counterfactual explanation model, γ_1 and γ_2 are hyperparameters for balancing the three parts of the loss function. The above formula reveals the minimum change required for the input data to achieve the idealized result. By adjusting the FC values between abnormal brain regions of SZ patients, their state may be closer to normal. This method not only provides an intuitive explanation scheme, but also provides SZ patients and doctors with the guidance needed to treat the disease.

4 Experiments and results

4.1 Experimental setting

In this work, we use a support vector machine (SVM) classifier to perform the classification task on five SZ datasets. During the experiments, we evaluate the performance of different methods based on diagnostic accuracy (ACC $= \frac{TP+TN}{TP+TN+FP+FN}$), sensitivity (SEN $= \frac{TP}{TP+FN}$) and specificity (SPE $= \frac{TN}{TN+FP}$). FP, TP, FN, and TN represent false-positive, true-positive, false-negative, and true-negative classification results. To ensure fairness, all compared FS methods use SVM classifiers. The parameters of our method are set as $\alpha_1 = \alpha_2 = 0.4$, $\alpha_3 = 0.2$, $t_{max} = 100$, $\rho = 0.9$, $n = 8$, $L = 10$, $\gamma_1 = 0.5$ and $\gamma_2 = 1$. It is worth noting that we use a five-fold cross-validation strategy in all experiments.

4.2 Statistical analysis of FC features

In this set of experiments, we perform statistical analysis on the functional connectivity (FC) remaining after feature selection by our method to demonstrate the effectiveness of our method. For intuitiveness, we first show the FC features retained after feature selection by our method in Figure 3. As can be seen from Figure 3, there are 16 shared FCs in the five datasets, and these shared FCs are selected as features in different datasets, indicating that they are crucial in identifying SZ. In addition, these shared FCs are mainly distributed in key brain regions such as the prefrontal cortex (PFC), cingulate gyrus (CC), and hippocampus (HIP), which is consistent with the findings of existing studies on SZ in brain network abnormalities (Orellana and Slachevsky, 2013; Wei et al., 2021; Frankle et al., 2022; Haznedar et al., 2004).

We select the five most statistically significant FC values between SZ and NC based on the statistical significance of each dataset, and the results are shown in Figure 4. From Figure 4, we find that the FC values between SZ and NC show different distribution patterns in the five datasets. Specifically, in some datasets, the FC values of SZ patients are significantly higher than those of NC, while in other datasets, the FC values of SZ patients are

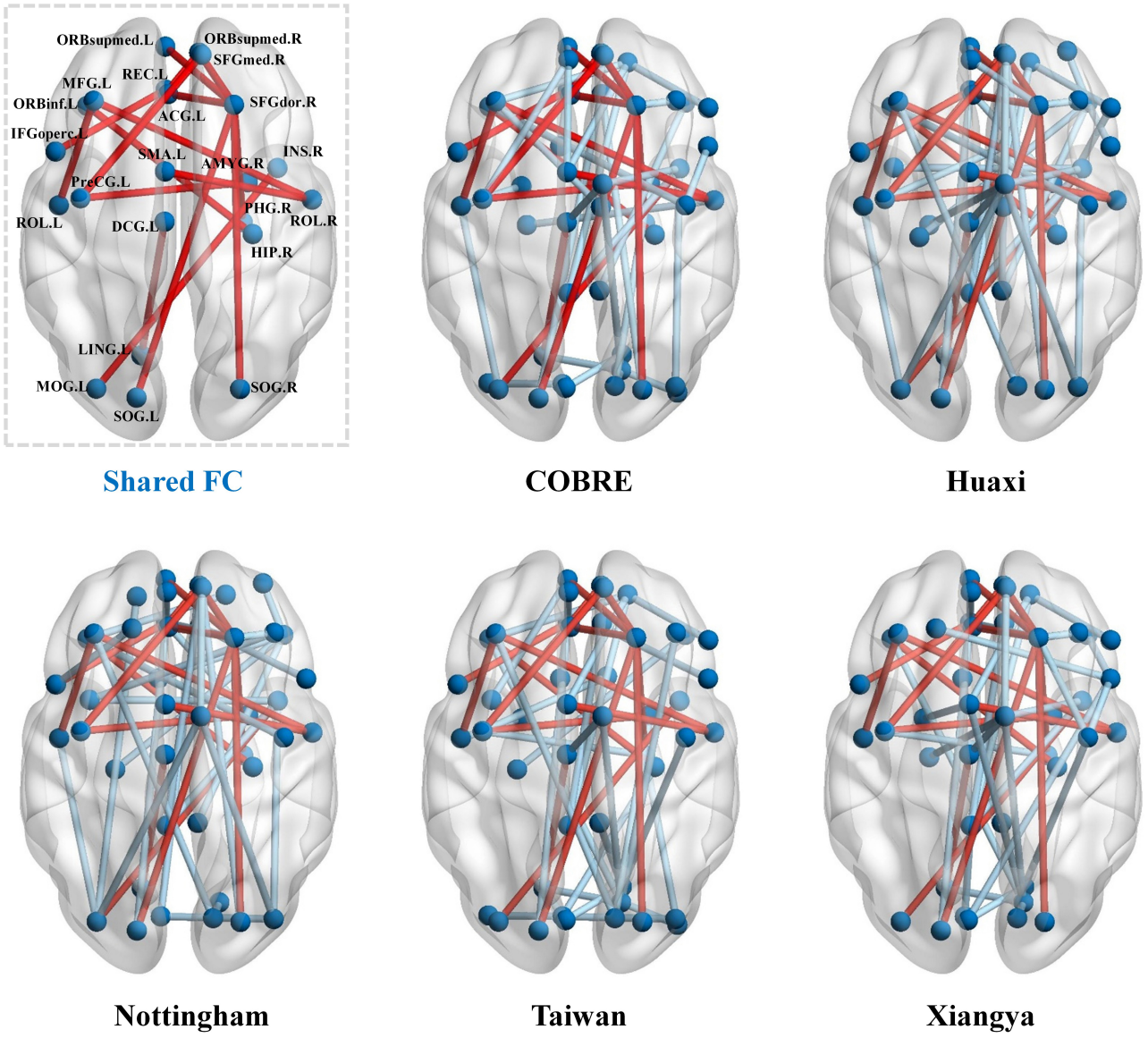

FIGURE 3
Functional connectivity (FC) retained after feature selection by our method in the five datasets. The red lines indicate the common functional connectivity among the five datasets.

significantly lower than those of NC. This suggests that there may be some heterogeneity in the functional connectivity patterns of SZ patients in different datasets. However, although the distribution of FC values in different datasets is different, some specific FCs show significant differences in multiple datasets, indicating that these FCs may play a key role in the neural mechanism of SZ.

Overall, the above results show that our method effectively extracts stable and biologically meaningful FC features, which helps to improve the accuracy and interpretability of SZ classification.

4.3 Comparison methods

We compare our proposed method with seven methods, including (i) RAW: classification without feature selection, as a baseline to illustrate the effect of applying feature selection techniques. (ii) LASSO: Lasso regression model based on L1 regularization (Cui et al., 2021). (iii) MFCSO: Multitasking Feature Selection via Competitive Swarm Optimizer (Li L. et al., 2023). (iv) MOEA\D: Multi-Objective Evolutionary Algorithm based on Decomposition (Wang et al., 2021). (v) SPEA: Strength Pareto Evolutionary Algorithm (Jiang and Yang, 2017). (vi) PSO-MET: Evolutionary Multitasking-Based Feature Selection via Particle Swarm Optimization (PSO) (Chen et al., 2020). (vii) MTPSO: Multitasking feature selection via PSO (Chen et al., 2021).

For all the above methods, the hyperparameters were set according to the values recommended in their respective original papers. Additionally, the number of iterations for all methods was set to 100, ensuring a consistent and fair comparison across all approaches.

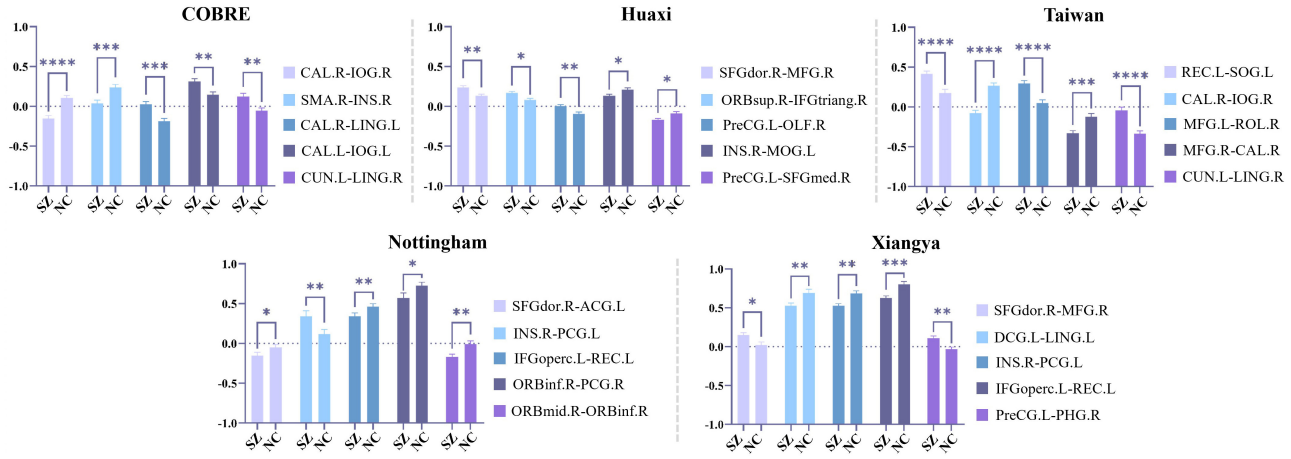

FIGURE 4
Statistical analysis is performed on the functional connections (FC) retained by feature selection using our method, and the five most statistically significant FC values between SZ and NC in the five datasets are shown here. Among them, * indicates $0.01 < p < 0.05$, ** indicates $0.001 < p < 0.01$, *** indicates $0.0001 < p < 0.001$, and **** indicates $p < 0.0001$.

MFCSO uses three filter methods for multi task feature selection, with each task optimized as an independent task without direct correlation between them. Therefore, the feature selection process may lack consistency. When dealing with specific datasets, especially on the schizophrenia (SZ) dataset, MFCSO may not be able to ensure consistency of selected features across different tasks, which may result in unstable performance on different datasets. Due to the lack of inter task correlation, feature selection results may be affected by randomness, making it difficult to effectively capture stable features related to schizophrenia.

Multi-objective evolutionary algorithms, such as MOEA\D and SPEA, are designed to address multiple objectives in feature selection. These algorithms provide a better balance between accuracy and feature diversity by considering multiple criteria in the optimization process. However, they are computationally intensive and can be prone to converging to local optima, especially in high-dimensional spaces. Furthermore, they often struggle with the trade-off between model complexity and accuracy, which can result in overfitting in small-sample scenarios, limiting their generalization ability.

PSO-MET and MTPSO are both particle swarm optimization-based methods that aim to improve feature selection by leveraging the concept of multitasking. While these methods are effective at identifying relevant features in some cases, they tend to be overly sensitive to initial conditions and parameter settings, leading to performance fluctuations. The lack of consistency across tasks and datasets reduces their reliability, particularly in real-world clinical settings where the data may be noisy or heterogeneous.

In comparison, our proposed method integrates robust multi-task feature selection with counterfactual explanation, offering several advantages over the methods discussed above. By using the Gray Wolf Optimizer (GWO) for feature selection, we ensure that our method not only handles high-dimensional data efficiently but also maintains stability across different datasets. The multi-task learning framework in our method allows for the sharing of knowledge across tasks, which improves generalization and reduces the risk of overfitting, particularly in small-sample situations.

4.4 Parameter analysis

In this section, we investigate the impact of varying the number of tasks on the performance of our multi-task optimization framework, as shown in the Figure 5. We observe that increasing the number of tasks generally leads to improvements in classification accuracy, especially for datasets such as Taiwan and Xiangya. These datasets achieve their highest classification accuracy at around six–nine tasks, where the accuracy reaches 0.87 and 0.89, respectively. This indicates that knowledge sharing between tasks is particularly effective in enhancing model performance when the task number is moderate. However, beyond a certain point, specifically around 10–12 tasks, the performance begins to plateau, with only marginal improvements in classification accuracy. The graph clearly shows that the datasets, such as Xiangya and Nottingham, while still improving with increasing task numbers, experience diminishing returns as the number of tasks exceeds 10. This suggests that while task number does play a role in boosting performance, there is an optimal task count that provides the best trade-off between performance enhancement and computational cost.

A deeper analysis reveals that the knowledge sharing between tasks is highly beneficial for improving classification performance. As the number of tasks increases, the model can leverage a broader range of features, which enhances its ability to generalize. However, once the number of tasks exceeds a threshold, redundancy starts to creep into the shared knowledge. This results in the transmission of features that do not contribute significantly to the performance improvement, thereby leading to a less efficient model. The redundancy of features becomes particularly evident when the number of tasks increases beyond 10, where the

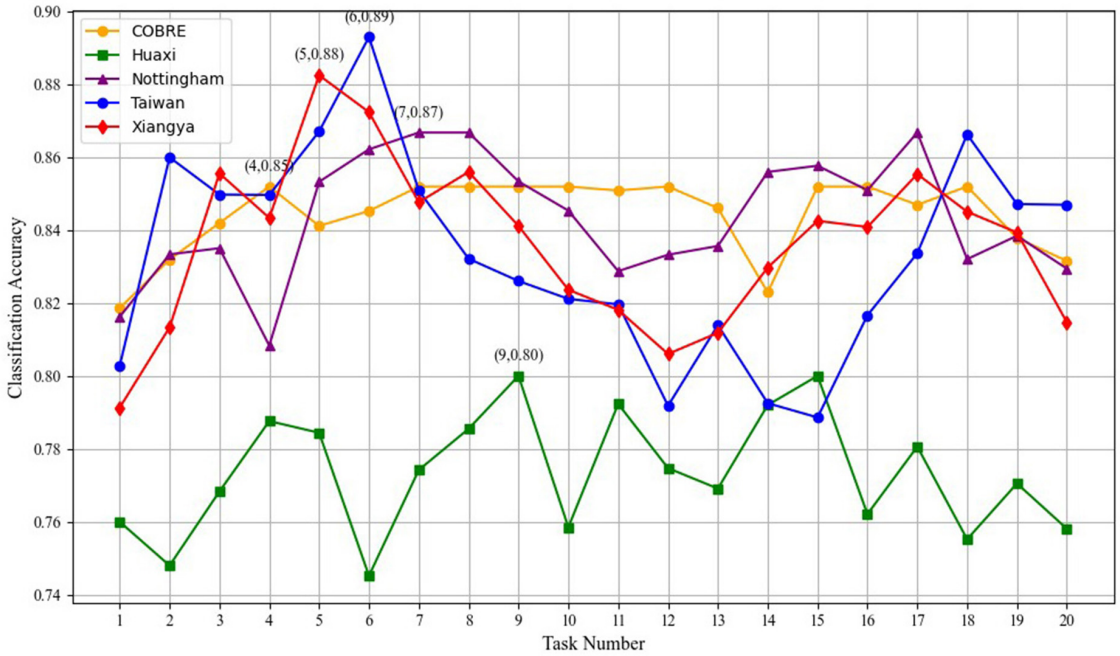

FIGURE 5
Impact of varying task numbers on model performance.

TABLE 2 Classification performance comparison with existing methods.

Datasets	Metric	RAW	LASSO	MFCSO	MOEA\D	SPEA	PSO-MET	MTPSO	Our method
COBRE	ACC (%)	63.41	75.00	68.10	73.40	69.44	78.38	81.19	**85.19**
	SEN (%)	58.33	66.67	60.00	76.47	78.57	68.75	**83.93**	80.00
	SPE (%)	73.68	79.17	76.19	69.77	63.64	85.71	79.45	**91.67**
Huaxi	ACC (%)	61.29	69.89	72.31	76.60	77.66	75.53	76.74	**80.00**
	SEN (%)	55.56	70.83	64.57	80.39	80.85	69.39	78.55	**82.86**
	SPE (%)	71.43	68.89	74.81	72.09	74.47	**82.22**	74.60	76.67
Nottingham	ACC (%)	65.00	66.12	72.22	72.34	75.53	80.95	82.71	**86.67**
	SEN (%)	66.67	68.97	66.67	74.51	76.60	80.00	82.13	**85.71**
	SPE (%)	63.64	62.50	77.78	69.77	74.47	81.82	83.08	**87.50**
Taiwan	ACC (%)	70.21	79.49	77.32	77.50	80.00	85.00	81.55	**89.29**
	SEN (%)	74.47	77.27	73.68	73.68	**88.24**	80.95	78.26	87.50
	SPE (%)	65.96	82.35	80.95	80.95	73.91	89.47	84.52	**91.67**
Xiangya	ACC (%)	66.90	69.23	79.41	76.74	70.77	72.31	82.79	**88.24**
	SEN (%)	51.35	58.82	72.22	72.00	74.29	68.57	**83.58**	80.00
	SPE (%)	67.39	69.73	87.50	83.33	66.67	76.67	81.79	**94.74**

Bold values represent the optimal values.

performance gains start to level off, and the computational overhead grows significantly.

Thus, while task quantity is crucial for leveraging task interdependencies and improving model accuracy, an excessive number of tasks may lead to inefficiency due to the sharing of redundant or less informative features. Therefore, it is essential to strike a balance between the number of tasks and the computational cost to ensure the model remains both effective and efficient.

4.5 Classification performance

In this set of experiments, we compare our proposed method with seven methods and show the results in Table 2. It is not difficult to see that our method shows excellent stability and consistency on the five datasets. Specifically, in the five datasets, the ACC of our method reaches 85.19% (COBRE), 80.00% (Huaxi), 86.67% (Nottingham), 89.29% (Taiwan), and 88.24% (Xiangya), while the ACC of most methods does not exceed 85%. Secondly, our method

performs outstandingly in both SEN and SPE, with SPE reaching 94.74% on the Xiangya dataset and SEN reaching 82.86% on the Huaxi dataset, indicating that our method has strong stability in the ability to distinguish between positive and negative samples. PSO-MET and MTPSO perform well in terms of SEN. For example, in the COBRE dataset, the SEN of MTPSO is 83.93%, which is higher than other methods, indicating that it has a strong ability to identify positive samples. In addition, we find that the methods based on multi-task optimization and evolutionary algorithms (i.e., PSO-MET and MTPSO) perform better overall. For example, in the Xiangya dataset, the ACC of MTPSO reaches 82.79%, which is significantly higher than other methods. This can be attributed to the fact that multi-task methods utilize shared knowledge across tasks, thereby improving the overall learning process. In general, the methods based on multi-task optimization and evolutionary algorithms have higher accuracy in SZ identification, while our method shows even better performance.

In addition, for the statistical significance of model performance, we select the three best-performing comparison methods (SPEA, PSO-MET, and MTPSO) in the experiment, and perform paired t-tests on the ACC indicators of each method on multiple datasets. The results are shown in Table 3. As can be seen from Table 3, our proposed method shows statistically significant differences with the three comparison methods on all datasets ($p < 0.05$). Specifically, the comparison with the SPEA method shows extremely significant differences on the COBRE, Nottingham, and Xiangya datasets ($p < 0.005$), and the comparison with PSO-MET has p values less than 0.025 on all datasets, indicating that the differences are highly statistically significant. At the same time, compared with the MTPSO method, although the p values in some datasets (such as Huaxi and COBRE) are relatively high, they do not exceed the significance level ($p < 0.05$), which still shows the stable advantages of our method on various datasets. These results further verify the universality and effectiveness of our method on multiple datasets from a statistical perspective.

4.6 Counterfactual explanations

In this set of experiments, we demonstrate how to generate a set of intuitive and diverse counterfactual (CF) examples for patients through the counterfactual explanation model. We provide counterfactual explanations by fine-tuning the abnormal FC value changes of patients, that is, adjusting the FC values between specific regions to make the patient's state closer to that of normal people. We generate two different counterfactual examples for SZ patients and present them in the form of brain maps and heat maps, as shown in Figure 6. It is not difficult to see that we can make the patient's state close to normal by only slightly adjusting the FC values between the corresponding regions. Specifically, in the Huaxi dataset, CF1 increases the FC values between ORBinf.R–HIP.L, SMA.R–SFGmed.R, SFGmed.L–ORBsupmed.L, and SMA.R–PHG.L from –0.2994, 0.0043, 0.2313, and 0.6822 to 0.1712, 0.8632, 0.2981, and 1.2072, and decreases the FC values between MFG.L–ROL.R and SFGdor.R–MFG.R from 0.1875 and 0.4143 to –0.6375 and –0.4230. In the Xiangya dataset, CF1 decreases the FC values between MFG.L–ROL.R, SFGdor.R–SOG.R, SFGdor.R–ACG.L, ORBsup.R–IFGtriang.R, and CUN.L–LING.R from 0.2149, 0.0883, –0.0146, –0.3282, and –0.0603 to –0.5490, 0.0619, –0.4669, –0.4412, and –0.8791, and increases the FC values between ORBsup.R–PCG.L and INS.R–PCG.L from –0.1435 and 0.4575 to 0.6884 and 1.2428, respectively. We find that the changes in functional connectivity (FC) after counterfactual interpretation remain stable within 1, without large-scale fluctuations, which further illustrates the robustness of our method. In addition, the role of FC changes in SZ patients has been observed in a large number of studies, such as Lynall et al. (2010), Fornito and Bullmore (2015), and Li et al. (2017).

TABLE 3 The t-test p-value results of our method and the three best performing comparison methods (SPEA, PSO-MET and MTPSO) on ACC.

Datasets	SPEA/our	PSO-MET/our	MTPSO/our
COBRE	0.0015	0.0220	0.0490
Huaxi	0.0439	0.0133	0.0269
Nottingham	0.0037	0.0019	0.0249
Taiwan	0.0143	0.0195	0.0174
Xiangya	0.0016	0.0029	0.0428

5 Discussion

In this paper, we propose a multi-task feature selection method for SZ diagnosis, and combine it with the counterfactual explanation model to fine-tune the abnormal FC features of SZ patients to make their state closer to that of healthy individuals, thereby improving the accuracy of SZ classification and the interpretability of the model. To demonstrate the effectiveness of our method, we conduct empirical studies on five SZ datasets. Our results show that across the five datasets, 16 FC features are selected simultaneously. These shared FC features are mainly distributed in key brain regions such as the prefrontal cortex (PFC), cingulate gyrus (CC) and hippocampus (HIP), which are widely considered to be closely related to the pathological mechanism of SZ in previous studies. For example, the study by Minzenberg et al. (2009) shows that PFC dysfunction is closely related to executive function deficits in SZ patients. Whitfield-Gabrieli et al. (2009) find that SZ patients have significant abnormalities in FC in the default mode network (including CC), which is associated with cognitive dysfunction. Gangadin et al. (2021) and Li X.-W. et al. (2023) find that SZ patients have significant abnormalities in FC between HIP and other brain regions in the resting state. These results not only verify that the abnormal FC features screened out by our method under multiple datasets are consistent and stable, but also further confirm its potential value in the diagnosis and interpretation of SZ from a neurobiological perspective.

Although previous studies reveal a variety of brain FC abnormalities associated with SZ, there is still a lack of an interpretable diagnostic tool in the diagnosis of SZ. Our study proposes an innovative method that integrates multi-task feature selection and counterfactual explanation. To generate accurate counterfactual examples, we construct a

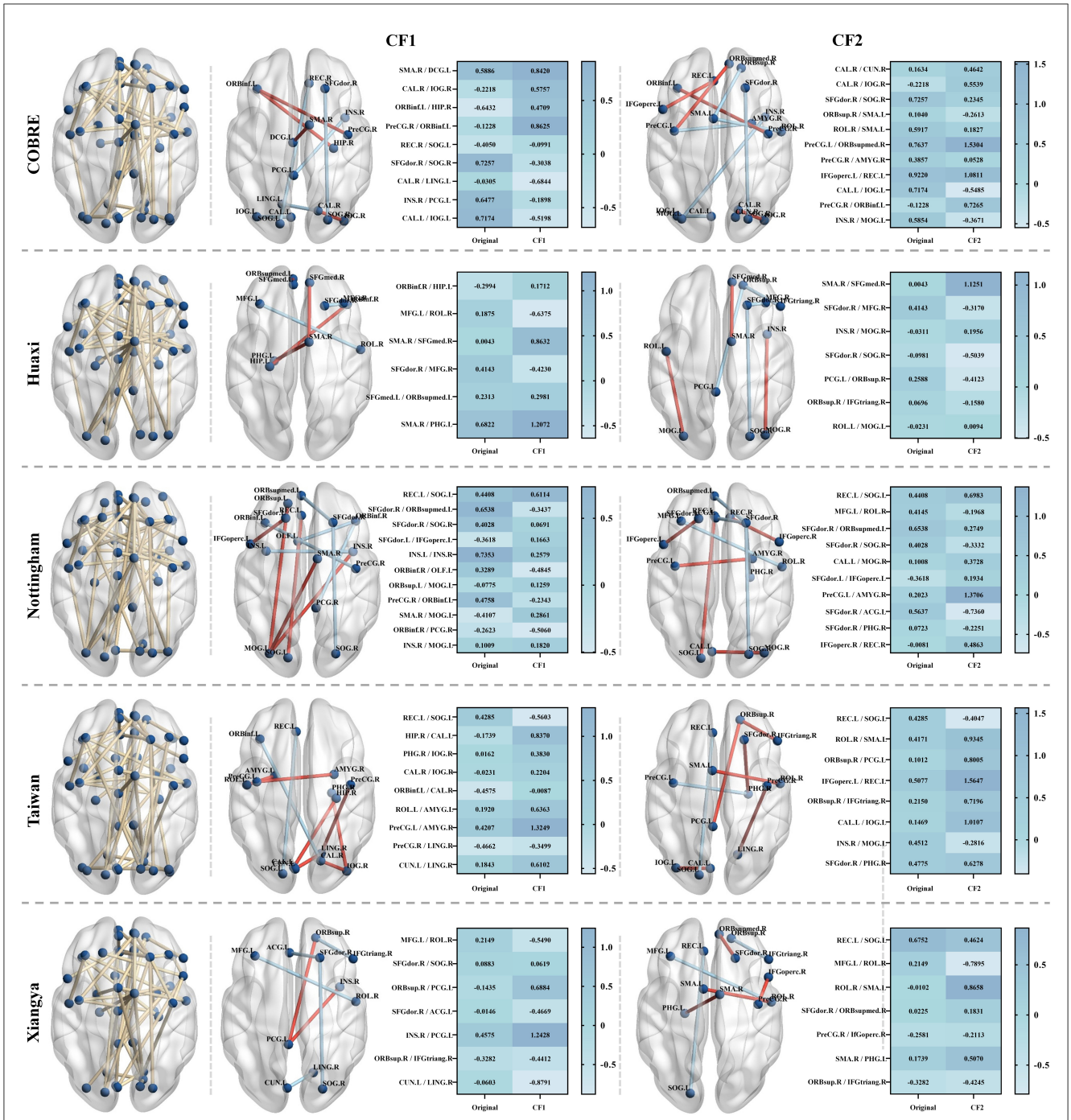

FIGURE 6
Counterfactual (CF) examples generated for randomly selected SZ patients in the test set in the five datasets. The yellow lines in the left brain map indicate the remaining functional connectivity (FC) after feature selection by our method, and the blue nodes indicate the corresponding brain regions. The brain maps in CF1 and CF2 show two counterfactual examples generated for the abnormal FC of the patients. The red lines indicate an increase in the FC value between the corresponding regions, and the blue lines indicate a decrease in the FC value. The heat map on the right shows the original FC value between the corresponding brain regions of the patients and the FC value after the counterfactual explanation.

counterfactual explanation model through three parts: loss function *loss*(·), distance function *dist*(·), and diversity index *diversity*(·). Specifically, *loss*(·) pushes counterfactual examples toward different predictions, *dist*(·) brings the counterfactual example closer to the original input, and *diversity*(·) increases the diversity of counterfactual explanations. We capture the brain regions where patients show abnormal FC features and slightly adjust the FC values between abnormal brain regions to make them closer to the normal state. This analysis method not only improves the interpretability of the classification model, but also provides an intuitive individual-level explanatory perspective for understanding brain FC abnormalities in SZ

patients, which helps to identify potential intervention targets and promotes the application of precision medicine in the diagnosis of SZ.

However, the current study still has several limitations. First, we only use the AAL model to define brain regions. In the future, we use different templates to evaluate the effectiveness of our proposed method. Second, we have not yet established cooperation with clinical medical institutions and lack counterfactual change explanations reviewed by clinicians. We plan to introduce clinical validation to further demonstrate the practicality and effectiveness of the method. Finally, this study focuses on the SZ dataset and further verifies the generalization ability and application potential of the method on other brain disease datasets such as Alzheimer's disease and autism.

6 Conclusion

In this paper, we propose a robust feature selection method based on multi-task optimization for SZ identification, and explain the changes in brain functional connectivity caused by the disease through a counterfactual explanation model. Compared with traditional methods, our proposed method not only improves the recognition performance, but also provides an intuitive explanation for the prediction of SZ, and verifies the effectiveness of the method on five SZ datasets.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

XY: Investigation, Methodology, Writing – original draft, Writing – review & editing. SW: Data curation, Investigation, Writing – original draft. YS: Methodology, Writing – review & editing. LG: Validation, Writing – review & editing. YH: Formal analysis, Writing – review & editing. TC: Formal analysis, Writing – review & editing. HY: Resources, Software, Writing – review & editing. HR: Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We sincerely appreciate the researchers and institutions that provided the publicly available datasets used in this study, including COBRE, Huaxi, Nottingham, Taiwan and Xiangya. These datasets have greatly contributed to the advancement of schizophrenia research. Additionally, we acknowledge the efforts of all participants and staff involved in data collection and preprocessing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abrate, C., and Bonchi, F. (2021). "Counterfactual graphs for explainable classification of brain networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (New York, NY: ACM), 2495–2504. doi: 10.1145/3447548.3467154

Chan, Y. H., Girish, D., Gupta, S., Xia, J., Kasi, C., He, Y., et al. (2024). Discovering robust biomarkers of psychiatric disorders from resting-state functional MRI via graph neural networks: a systematic review. *arXiv* [Preprint]. arXiv:2405.00577. doi: 10.48550/arXiv.2405.00577

Chen, K., Xue, B., Zhang, M., and Zhou, F. (2020). An evolutionary multitasking-based feature selection method for high-dimensional classification. *IEEE Trans. Cybern.* 52, 7172–7186. doi: 10.1109/TCYB.2020.3042243

Chen, K., Xue, B., Zhang, M., and Zhou, F. (2021). Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimization. *IEEE Trans. Evol. Comput.* 26, 446–460. doi: 10.1109/TEVC.2021.3100056

Cheng, F., Ming, Y., and Qu, H. (2020). Dece: decision explorer with counterfactual explanations for machine learning models. *IEEE Trans. Vis. Comput. Graph.* 27, 1438–1447. doi: 10.1109/TVCG.2020.3030342

Chyzhyk, D., Savio, A., and Graña, M. (2015). Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of elm. *Neural Netw*. 68, 23–33. doi: 10.1016/j.neunet.2015.04.002

Cui, L., Bai, L., Wang, Y., Yu, P. S., and Hancock, E. R. (2021). Fused lasso for feature selection using structural information. *Pattern Recognit*. 119:108058. doi: 10.1016/j.patcog.2021.108058

Ding, W., Zhou, T., Huang, J., Jiang, S., Hou, T., Lin, C.-T., et al. (2024). FMDNN: a fuzzy-guided multi-granular deep neural network for histopathological image classification. *IEEE Trans. Fuzzy Syst*. 32, 4709–4723. doi: 10.1109/TFUZZ.2024.3410929

Fornito, A., and Bullmore, E. T. (2015). Reconciling abnormalities of brain network structure and function in schizophrenia. *Curr. Opin. Neurobiol*. 30, 44–50. doi: 10.1016/j.conb.2014.08.006

Frankle, W. G., Himes, M., Mason, N. S., Mathis, C. A., and Narendran, R. (2022). Prefrontal and striatal dopamine release are inversely correlated in schizophrenia. *Biol. Psychiatry* 92, 791–799. doi: 10.1016/j.biopsych.2022.05.009

Gangadin, S. S., Cahn, W., Scheewe, T. W., Pol, H. E. H., and Bossong, M. G. (2021). Reduced resting state functional connectivity in the hippocampus-midbrain-striatum network of schizophrenia patients. *J. Psychiatr. Res*. 138, 83–88. doi: 10.1016/j.jpsychires.2021.03.041

Haznedar, M. M., Buchsbaum, M. S., Hazlett, E. A., Shihabuddin, L., New, A., Siever, L. J., et al. (2004). Cingulate gyrus volume and metabolism in the schizophrenia spectrum. *Schizophr. Res*. 71, 249–262. doi: 10.1016/j.schres.2004.02.025

Hu, R., Peng, Z., Zhu, X., Gan, J., Zhu, Y., Ma, J., et al. (2021). Multi-band brain network analysis for functional neuroimaging biomarker identification. *IEEE Trans. Med. Imaging* 40, 3843–3855. doi: 10.1109/TMI.2021.3099641

Huang, J., Wang, M., Ju, H., Ding, W., and Zhang, D. (2025). Agbn-transformer: anatomy-guided brain network transformer for schizophrenia diagnosis. *Biomed. Signal Process. Control* 102:107226. doi: 10.1016/j.bspc.2024.107226

Insel, T. R. (2010). Rethinking schizophrenia. *Nature* 468, 187–193. doi: 10.1038/nature09552

Jiang, S., and Yang, S. (2017). A strength pareto evolutionary algorithm based on reference direction for multiobjective and many-objective optimization. *IEEE Trans. Evol. Comput*. 21, 329–346. doi: 10.1109/TEVC.2016.2592479

Li, L., Xuan, M., Lin, Q., Jiang, M., Ming, Z., Tan, K. C., et al. (2023). An evolutionary multitasking algorithm with multiple filtering for high-dimensional feature selection. *IEEE Trans. Evol. Comput*. 27, 802–816. doi: 10.1109/TEVC.2023.3254155

Li, T., Wang, Q., Zhang, J., Rolls, E. T., Yang, W., Palaniyappan, L., et al. (2017). Brain-wide analysis of functional connectivity in first-episode and chronic stages of schizophrenia. *Schizophr. Bull*. 43, 436–448. doi: 10.1093/schbul/sbw099

Li, X.-W., Liu, H., Deng, Y.-Y., Li, Z.-Y., Jiang, Y.-H., Li, D.-Y., et al. (2023). Aberrant intra-and internetwork functional connectivity patterns of the anterior and posterior hippocampal networks in schizophrenia. *CNS Neurosci. Ther*. 29, 2223–2235. doi: 10.1111/cns.14171

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., et al. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Med. Image Anal*. 20, 112–134. doi: 10.1016/j.media.2014.10.011

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci*. 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Matsui, T., Taki, M., Pham, T. Q., Chikazoe, J., and Jimura, K. (2022). Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network. *Front. Neuroinform*. 15:802938. doi: 10.3389/fninf.2021.802938

McCutcheon, R. A., Marques, T. R., and Howes, O. D. (2020). Schizophrenia—an overview. *JAMA Psychiatry* 77, 201–210. doi: 10.1001/jamapsychiatry.2019.3360

Mhiri, I., and Rekik, I. (2020). Joint functional brain network atlas estimation and feature selection for neurological disorder diagnosis with application to autism. *Med. Image Anal*. 60:101596. doi: 10.1016/j.media.2019.101596

Minzenberg, M. J., Laird, A. R., Thelen, S., Carter, C. S., and Glahn, D. C. (2009). Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch. Gen. Psychiatry* 66, 811–822. doi: 10.1001/archgenpsychiatry.2009.91

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Adv. Eng. Softw*. 69, 46–61. doi: 10.1016/j.advengsoft.2013.12.007

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency* (New York, NY: ACM), 607–617. doi: 10.1145/3351095.3372850

Naheed, N., Shaheen, M., Khan, S. A., Alawairdhi, M., and Khan, M. A. (2020). Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. *Comput. Model. Eng. Sci*. 125, 314–344. doi: 10.32604/cmes.2020.011380

Orellana, G., and Slachevsky, A. (2013). Executive functioning in schizophrenia. *Front. Psychiatry* 4:35. doi: 10.3389/fpsyt.2013.00035

Prado-Romero, M. A., Prenkaj, B., Stilo, G., and Giannotti, F. (2023). A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. *ACM Comput. Surv*. 56, 1–37. doi: 10.1145/3618105

Rantala, M. J., Luoto, S., Borráz-León, J. I., and Krams, I. (2022). Schizophrenia: the new etiological synthesis. *Neurosci. Biobehav. Rev*. 142:104894. doi: 10.1016/j.neubiorev.2022.104894

Richens, J. G., Lee, C. M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun*. 11:3923. doi: 10.1038/s41467-020-17419-7

Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., and Cristani, M. (2020). Infinite feature selection: a graph-based feature filtering approach. *IEEE Trans. Pattern Anal. Mach. Intell*. 43, 4396–4410. doi: 10.1109/TPAMI.2020.3002843

Song, X., Wu, K., and Chai, L. (2023). Brain network analysis of schizophrenia patients based on hypergraph signal processing. *IEEE Trans. Image Process*. 32, 4964–4976. doi: 10.1109/TIP.2023.3307975

Spreitzer, N., Haned, H., and van der Linden, I. (2022). "Evaluating the practicality of counterfactual explanations," in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS* 2022.

Sunil, G., Gowtham, S., Bose, A., Harish, S., and Srinivasa, G. (2024). Graph neural network and machine learning analysis of functional neuroimaging for understanding schizophrenia. *BMC Neurosci*. 25:2. doi: 10.1186/s12868-023-00841-0

Turner, B. O., Paul, E. J., Miller, M. B., and Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol*. 1:62. doi: 10.1038/s42003-018-0073-z

Verma, S., Goel, T., Tanveer, M., Ding, W., Sharma, R., Murugan, R., et al. (2023). Machine learning techniques for the schizophrenia diagnosis: a comprehensive review and future research directions. *J. Ambient Intell. Humaniz. Comput*. 14, 4795–4807. doi: 10.1007/s12652-023-04536-6

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. JL Tech*. 31:841. doi: 10.2139/ssrn.3063289

Wang, P., Xue, B., Liang, J., and Zhang, M. (2021). Multiobjective differential evolution for feature selection in classification. *IEEE Trans. Cybern*. 53, 4579–4593. doi: 10.1109/TCYB.2021.3128540

Wang, Y., Li, Z., Wang, Y., Wang, X., Zheng, J., Duan, X., et al. (2015). A novel approach for stable selection of informative redundant features from high dimensional fMRI data. *arXiv* [Preprint]. arXiv:1506.08301. doi: 10.48550/arXiv.1506.08301

Wei, G.-X., Ge, L., Chen, L.-Z., Cao, B., and Zhang, X. (2021). Structural abnormalities of cingulate cortex in patients with first-episode drug-naïve schizophrenia comorbid with depressive symptoms. *Hum. Brain Mapp*. 42, 1617–1625. doi: 10.1002/hbm.25315

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., et al. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proc. Nat. Acad. Sci*. 106, 1279–1284. doi: 10.1073/pnas.0809141106

Xing, Y., Kochunov, P., van Erp, T. G., Ma, T., Calhoun, V. D., Du, Y., et al. (2022). A novel neighborhood rough set-based feature selection method and its application to biomarker identification of schizophrenia. *IEEE J. Biomed. Health Inform*. 27, 215–226. doi: 10.1109/JBHI.2022.3212479

Zhang, X., Braun, U., Harneit, A., Zang, Z., Geiger, L. S., Betzel, R. F., et al. (2021). Generative network models of altered structural brain connectivity in schizophrenia. *Neuroimage* 225:117510. doi: 10.1016/j.neuroimage.2020.117510

Zhu, C., Tan, Y., Yang, S., Miao, J., Zhu, J., Huang, H., et al. (2024). Temporal dynamic synchronous functional brain network for schizophrenia classification and lateralization analysis. *IEEE Trans. Med. Imaging* 43, 4307–4318. doi: 10.1109/TMI.2024.3419041