scientific reports

OPEN

Counterfactual explanations of tree based ensemble models for brain disease analysis with structure function coupling

Shaolong Wei[1,2], Zhen Gao[2]✉, Hongcheng Yao[3], Xiaoyu Qi[1], Mingliang Wang[4]✉ & Jiashuang Huang[1]✉

Convergent evidence has suggested that the disruption of either structural connectivity (SC) or functional connectivity (FC) in the brain can lead to various neuropsychiatric disorders. Since changes in SC-FC coupling may be more sensitive than a single modality to detect subtle brain connectivity abnormalities, a few learning-based methods have been proposed to explore the relationship between SC and FC. However, these existing methods still fail to explain the relationship between altered SC-FC coupling and brain disorders. Therefore, in this paper, we explore three types of tree-based ensemble models (i.e., Decision Tree, Random Forest, and Adaptive Boosting) toward counterfactual explanations for SC-FC coupling. Specifically, we first construct SC and FC matrices from preprocessed diffusion-weighted DTI and resting-state functional fMRI data. Then, we quantify the SC-FC coupling strength of each region and convert it into feature vectors. Subsequently, we select SC-FC coupling features that can reflect disease-related information and trained three tree-based models to analyze the predictive role of these coupling features for diseases. Finally, we design a tree ensemble counterfactual explanation model to generate a set of counterfactual examples for patients, thereby assisting the diagnosis of brain diseases by fine-tuning the patient's abnormal SC-FC coupling feature vector. Experimental results on two independent datasets (i.e., epilepsy and schizophrenia) validate the effectiveness of the proposed method. The identified discriminative brain regions and generated counterfactual examples provide new insights for brain disease analysis.

Keywords Structural connectivity, Functional connectivity, SC-FC coupling, Brain diseases, Counterfactual explanation

Our brain is a complex neural network composed of numerous interconnected structures and functional regions[1–3]. The structural connectivity (SC) of the brain, in conjunction with other factors, shapes neurophysiological activity and thus affects the functional connectivity (FC) between neuronal populations and brain regions[4]. SC reflects the connections between physical white matter regions in the brain, which can be obtained through diffusion tensor imaging (DTI), while FC describes the temporal co-activation of activity between these regions, typically constructed by functional magnetic resonance imaging (fMRI)[5–7]. Multiple studies have shown that brain diseases can disrupt normal brain networks[8–12]. For example, in many brain diseases such as epilepsy, schizophrenia, and brain tumors, a series of abnormal connectivity or network characteristics have been found[13–19]. Therefore, analyzing large-scale brain networks helps to reveal the pathological and physiological mechanisms of brain diseases.

The relationship between structural connectivity and functional connectivity (known as SC-FC coupling) has been studied at the overall level of brain diseases[20,21]. SC-FC coupling describes the functional dynamics of the brain from the perspective of structural topology, and more sensitively detects subtle brain changes caused by diseases than any single imaging method[22,23]. Currently, some learning-based methods have been proposed to explore the relationship between SC and FC. For example, Popp et al.[24] used a network communication model to capture SC-FC coupling in different regions of the brain, thereby proving that SC-FC coupling is related

[1]School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China. [2]Affiliated Hospital 2 of Nantong University, Nantong, China. [3]School of Information Science and Technology, Nantong University, Nantong, China. [4]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. ✉email: 87801923@163.com; wml489@nuist.edu.cn; hjshdym@163.com

to human cognitive ability. Dan et al.[25] proposed a physically guided deep model to reveal the mechanistic role of SC-FC coupling changes in disease progression. Sarwar et al.[26]trained a deep neural network to predict brain functional connectivity based on structural connectivity groups. In addition, SC-FC coupling plays an important role in various brain diseases and is currently a key biomarker for diagnosing and predicting disease progression[27]. Gu et al.[28]'s work has demonstrated that regional SC-FC coupling is a special feature of brain tissue and is influenced by genetic factors. Liao et al.[29] proposed that SC-FC coupling is not only influenced by diseases, but also related to clinical manifestations. The research results of Chen et al.[30] have shown that SC-FC coupling may serve as a valuable characteristic biomarker for the burden of Parkinson's disease. Kong et al.[31] found that SC-FC coupling exhibits significant changes in different stages of schizophrenia.

In the existing analyses of SC-FC coupling, various models have been employed to explore the relationship between brain structure and function. For example, statistical models evaluate SC-FC coupling by calculating the correlation between SC and FC[32], biophysical models simulate the relationship between SC and FC based on biological mechanisms[33], and network communication models use specific strategies based on neural communication to understand the SC-FC coupling mechanism[34]. In addition, computational models study the dynamic interactions between SC and FC through simulation and optimization methods[35], while connectomics models focus on revealing the topological structure and functional connectivity patterns between different brain regions to gain a deeper understanding of the basis of SC and FC coupling[36]. Although these models reveal the complex relationship between SC and FC from different perspectives, there is still a lack of an interpretable auxiliary diagnostic tool. Counterfactual explanation provides an actionable and humane way of explanation to address this issue[37,38]. Cheng et al.[39] introduced a classic example of counterfactuals: a person submitted a loan application but was rejected by the bank. If his credit score were 700 instead of 600, his loan application would have been approved.

Counterfactual explanations have been widely used in the field of disease diagnosis. For example, Singla et al.[40] provided a BlackBox counterfactual interpreter aimed at explaining image classification models for medical applications. Richens et al.[41] rephrased medical diagnosis as a counterfactual reasoning task by establishing a counterfactual causal diagnosis model, improving the effectiveness of machine learning in the field of medical diagnosis. Kim et al.[42]proposed a new generative network based on counterfactual explanation, which can effectively detect lesions. Currently, most existing counterfactual explanation methods usually rely on the differentiability of the model, which may not apply to non-differentiable models such as tree ensembles because they do not use gradient-based optimization procedures. However, compared with traditional counterfactual explanation models, the tree ensemble model can identify the most important features for brain disease diagnosis, which helps to generate more accurate counterfactual explanations[43–45].

In this paper, we propose a novel brain disease analysis method that aims to reveal the potential impact of the coupling between SC and FC on brain diseases through counterfactual explanations. Unlike existing methods, we combine three tree-based ensemble models (i.e. Decision Tree, Random Forest, and Adaptive Boosting) to analyze and explain SC-FC coupling features and construct counterfactual examples to provide personalized explanations for patients. The schematic diagram of our proposed method is shown in Figure 1. Specifically, we first construct SC and FC matrices from preprocessed diffusion-weighted DTI and resting-state functional fMRI data. Then, we quantify the SC-FC coupling strength of each region and convert it into feature vectors. Subsequently, we select SC-FC coupling features that can reflect disease-related information and trained three tree-based models to analyze the predictive role of these coupling features for diseases. Finally, we design a tree ensemble counterfactual explanation model by fine-tuning the abnormal SC-FC coupling features of patients to generate a set of counterfactual examples that help us understand which feature changes have a significant impact on the predicted outcomes. We validate our proposed method on epilepsy and schizophrenia datasets, and the results show that our approach not only improved the interpretability of the model but also provided new insights into the analysis of brain diseases.

Materials and methods

Dataset

In this study, we use two datasets related to brain diseases: the epilepsy dataset and the schizophrenia dataset. It is worth noting that all subjects are right-handed. All patients in both datasets are diagnosed by experienced physicians to ensure the reliability and accuracy of the diagnostic results. All experiments are performed in accordance with relevant guidelines and regulations. All participants provide written informed consent prior to participation. The main details about the two datasets are as follows:

- **Epilepsy dataset:** The patient group consists of two types of epilepsy: 103 patients with frontal lobe epilepsy (FLE) and 89 patients with temporal lobe epilepsy (TLE). Recruit 114 healthy volunteers who are age and gender matched as controls. According to the Pearson χ^2 test, the gender p-value is 0.9601. The age p-value obtained from the one-way ANOVA test among the three groups is 0.2471. These results indicate that there are no significant differences in gender and age distribution among the groups. All participants are from Jinling Hospital of Nanjing University School of Medicine. The specific information of the participants is shown in Table 1.
- **Schizophrenia dataset:** The participants included 26 patients with schizophrenia (SZ) and 28 age and gender matched healthy volunteers. The p-value for gender is 0.74, and the p-value for age is 0.8023, indicating that there are no significant differences in gender and age distribution among the groups. All participants are from Nanjing Medical University. The specific information of the participants is shown in Table 2.

Data pre-processing

The detailed pre-processing process for epilepsy and schizophrenia datasets is described as follows:

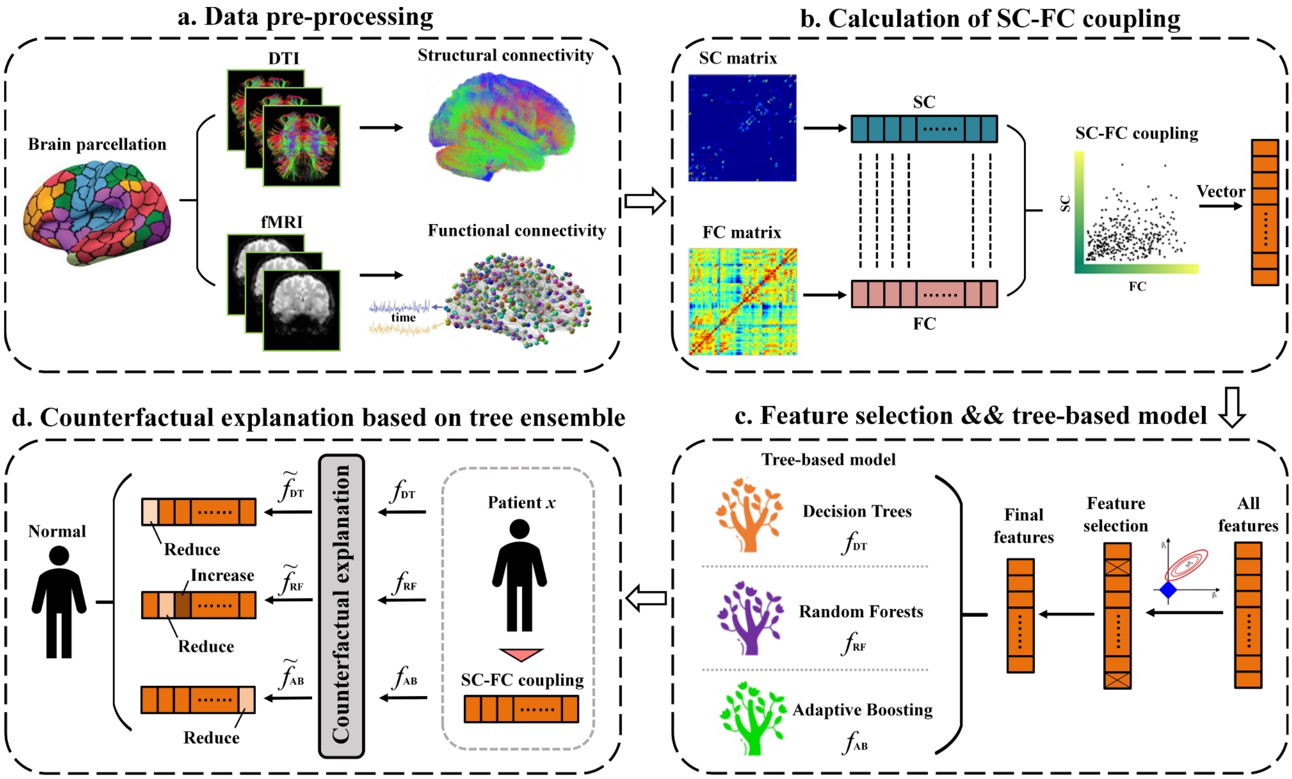

Fig. 1. Illustration of our proposed brain disease analysis method, including (**a**) data pre-processing, (**b**) calculation SC-FC coupling, (**c**) feature selection and tree-based model, (**d**) counterfactual explanation based on tree ensemble.

Group	FIE	TLE	Normal
No. of subjects(male/female)	53/50	44/45	58/56
Age range	17–51	17–51	20–38
Average age	24.1	25.9	26.2

Table 1. Epilepsy dataset.

Group	SZ	Normal
No. of subjects(male/female)	16/10	16/12
Age range	22–54	24–51
Average age	35.0	35.5

Table 2. Schizophrenia dataset.

- **Epilepsy dataset:** The image acquisition is performed by a Siemens Trio 3T scanner from Nanjing Jinling Hospital. Use foam pads to minimize head movement for all subjects. Use a single gradient recall echo plane imaging sequence to obtain functional scans. The scanning parameters are as follows: repetition time = 2000 *ms*; echo time = 30 *ms*; flip angle = 90°; 30 transverse slices; field of view (FOV) = 240×240 mm^2; slice thickness = 4 *ms*; gap = 0.4 *ms*; voxel size = 3.75×3.75×3.75 mm^3; obtain DTI scans using echo plane imaging sequences based on spin echoes. The scanning parameters are as follows: repetition time = 6100 *ms*; echo time = 93 *ms*; flip angle = 90°; field of view = 240×240 mm^2; matrix size = 256×256; voxel size = 0.94×0.94× 3 mm^3; 45 slices; In addition, data is collected using 30 optimal nonlinear diffusion gradient directions with b = 1000 s/mm^2 and a non-diffusion weighted volume. Repeat the entire sequence four times to improve the signal-to-noise ratio. The parameters for structural MRI T1 scan are as follows: repetition time = 2300 *ms*; echo time = 2.98 *ms*; flip angle = 90°; field of view = 240×240 mm^2; matrix size = 256×256, zero padding and interpolation to 512×512, slice thickness = 1 *mm*, no gap between slices, voxel size = 0.5×0.5×1 mm^3 and 176 slices. It is worth noting that MRI data is only used during the data processing phase.

- **Schizophrenia dataset:** The data is collected using a Siemens Trio 3T scanner. The scan parameters of rs-fMRI are as follows: TR = 2000 *ms*, TE = 21 *ms*, flipping angle = 90°, voxel size = 3.75×3.75×3.75 *mm*^{3}; The scan parameters of DTI are as follows: TR = 7400 *ms*, TE = 21 *ms*, flip angle = 90°, voxel size = 0.94×0.94×0.94 *mm*^{3}. Image pre-processing consistent with epilepsy dataset.

Brain network construction

We construct the structural network through DTI and the functional network through fMRI. Specifically, DTI data is processed using the PANDA suite. Firstly, use the FSL toolbox for DTI distortion correction (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki). This tool can remove eddy current and extract brain masks from non-diffusion weighted images (B0 images). Then, we use TrackVis to obtain fiber images based on the anatomical regions defined by the automated anatomical labeling (AAL) convention on the T1 images of each subject, and employ deterministic tracking methods. Finally, the number of fibers can naturally be seen as the strength of the edges in the structural brain network. The fMRI data is processed using spm12 and executed in the DPARSF Toolbox version 2.0[46]. The initial functional time series consists of slice time collection, correction, rearrangement, and normalization into EPI templates. Then, we perform trend removal to remove false sources of variance (i.e. six head motion parameters, average signals of cerebrospinal fluid and white matter, and global brain signals), and perform band-pass filtering on the time series (0.01–0.08Hz). The generated volume has 240 time points and is divided into 90 regions of interest (ROIs) using the AAL map[47]. Finally, the Pearson correlation between the entire network time processes is utilized to determine the weights of edges in the functional brain network.

Calculation of SC-FC coupling

We choose non-parametric Spearman correlation to quantify the coupling strength of structural and functional connectivity in a region, as it is an intuitive and easily interpretable measurement method. Importantly, it adapts to the non-Gaussian nature of entries in SC and FC[28]. As shown in Figure 1(b), SC-FC coupling is constructed by calculating the Spearman correlation coefficient between each row of the SC matrix and the corresponding row of the FC matrix. Analyzing the results of this step indicates that a vector with a length of 90 is obtained in each subject, representing the coupling strength of SC-FC in each of the 90 regions in the graph. We further construct the SC-FC feature matrix *X* for all subjects. Here, the dimension of *X* is $n \times 90$, where *n* represents the number of subjects. Each subject contains a vector with a dimension of 1×90, reflecting the subject's SC-FC coupling strength in 90 regions.

Feature selection and Tree-based model

To identify the most important SC-FC coupling features for brain disease diagnosis, we use the ***Lasso*** regression model for feature selection to remove unnecessary redundant information in matrix *X*[48]. After the above steps, we have a set of SC-FC coupling features $\{(\mathrm{x_i}, y_i), i = 1, 2, ..., n\}$, where *n* is the number of subjects. Each $\mathrm{x_i} = (x_i^1, x_i^2, ..., x_i^{90}) \in R^{1\times 90}$ is an SC-FC coupling vector, and $y_i \in R$ is its corresponding label for the i-th subject. We use matrix form $X = [\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x_n}]^T \in \mathbb{R}^{n\times 90}$ to represent the data and represent its corresponding labels as $y = (y_1, y_2, ..., y_n)^T \in \mathbb{R}^n$. With this representation method, the ***Lasso*** regression model can be written as:

$$\omega = \arg\min_\omega ||y - \omega X||_2^2 + \xi||\omega||_1 \tag{1}$$

Among them, ω is the coefficient vector and ξ is the regularization parameter. The selected SC-FC coupling feature matrix is represented as $X' \in \mathbb{R}^{n\times k}$, where $k \ll 90$.

Based on the selected feature matrix X', we train a series of machine learning models (i.e., $f : X' \to y$) for brain disease diagnosis. In this paper, we train three tree-based classification models, including Decision Tree (DT) (i.e., f_{DT}), Random Forest (RF) (i.e., f_{RF}), and Adaptive Boosting (AB) (i.e., f_{AB}). Among them, Decision trees serve as the basic learner, integrating multiple decision trees to build a powerful predictive model, thereby improving the accuracy of brain disease diagnosis.

Counterfactual explanation based on tree ensemble

After the above steps, we use a tree ensemble counterfactual explanation model to adjust the patient's SC-FC coupling features to make their state closer to that of a normal person. Specifically, the input of this model is the SC-FC coupling feature vector $\mathrm{x_i}' \in R^{1\times k}$ of the *i*-th subject and a tree-based classifier *f*. The necessary condition for using this counterfactual explanation model to generate the best counterfactual example is that the input model *f* must satisfy the differentiable condition. Obviously, none of the three models we trained (i.e. f_{DT}, f_{RF} and f_{AB}) satisfy the differentiable condition. Therefore, we adopt the method of Lucic et al.[49] to incorporate the differentiable approximation $\widetilde{f}$ of the non-differentiable model *f* into the gradient-based optimization framework. This method is achieved by replacing each partition in each tree with a *sigmoid* function, whose hyperparameter is η, defined as:

$$sig(u) = (1 + \exp(\eta \cdot u))^{-1} \tag{2}$$

where $\eta \in \mathbb{Z}_{>0}$. The *sigmoid* function is incorporated into function $\widetilde{m_q}\ (\mathrm{x_i}')$, which is the $m_q(\mathrm{x_i}')$ function to approximates the activation of node *q*: $m_q(\mathrm{x_i}') \approx \widetilde{m_q}\ (\mathrm{x_i}')$. The function is defined as follows:

$$\widetilde{m_q}\left(\mathrm{x_i}'\right) = \begin{cases} 1, & \text{if } q \text{ is the root,} \\ \widetilde{m_Q}\left(\mathrm{x_i}'\right) \cdot sig(\theta_q - x_i^j), & \text{if } q \text{ is a left child,} \\ \widetilde{m_Q}\left(\mathrm{x_i}'\right) \cdot sig(x_i^j - \theta_q), & \text{if } q \text{ is a right child.} \end{cases} \tag{3}$$

where θ_q is a threshold for activation of node q, which is determined by the model's determination of the optimal segmentation point during the training phase. For ease of explanation, we present a simple decision tree T and its corresponding differentiable approximation tree $\widetilde{T}$ in Figure 2. If the parent node Q of node q is activated and the j-th SC-FC coupling feature $x_i^j (j \leq k) > \theta_q$ of the i-th subject, then node q (left child node) is activated. If $x_i^j < \theta_q$, then node $\widehat{q}$ (right child node) is activated. The root node is always activated. Then, by approximating the activation of node q through the *sigmoid* function, the differentiable approximation tree $\widetilde{T}$ of a single decision tree T can be obtained. So the tree approximation can be defined as:

$$\widetilde{T}\left(y_i|\mathrm{x_i}'\right) = \sum_{q \in \Omega_T} \widetilde{m_q}\left(\mathrm{x_i}'\right) \cdot T(y_i|q) \tag{4}$$

Among them, Ω_T is the set of leaf nodes in T, and each leaf node q has its own predicted distribution $T(y_i|q)$. From Figure 2, it can be seen that the approximate tree $\widetilde{T}$ and T has the same tree structure and threshold, but compared to tree T, its node activation mainly depends on the feature x_i^j and threshold θ_q. Finally, this method replaces the maximum value operation of f with *softmax* function of temperature $\lambda \in \mathbb{Z}_{>0}$, where f is an ensemble of L trees with weights $\rho_l \in \mathbb{Z}$. The approximate value $\widetilde{f}$ can be expressed as:

$$\widetilde{f}\left(y_i|\mathrm{x_i}'\right) = \frac{\exp(\lambda \cdot \sum_{l=1}^{L} \rho_l \cdot \widetilde{T_l}\left(y_i|\mathrm{x_i}'\right))}{\sum_{y_i'} \exp(\lambda \cdot \sum_{l=1}^{L} \rho_l \cdot \widetilde{T_l}\left(y_i'|\mathrm{x_i}'\right))} \tag{5}$$

where $y_i{}'$ is the predicted label. In our experiment, 0 corresponds to normal subjects and 1 corresponds to patients. We want to adjust the SC-FC coupling features of patients to make them closer to normal, so we usually set $y_i{}'$ to 0. After obtaining the approximate value $\widetilde{f}$ based on the original model f, we can obtain the optimal counterfactual example $\overset{*}{\mathrm{x_i}}{}'$:

$$\overset{*}{\mathrm{x_i}}{}' = \arg\min_{\overline{\mathrm{x_i}}'} \widetilde{f}\left(y_i{}'|\mathrm{x_i}'\right) + \alpha \cdot d(\mathrm{x_i}', \overline{\mathrm{x_i}}') \tag{6}$$

where weights $\alpha \in \mathbb{Z}_{>0}$, $\overline{\mathrm{x_i}}'$ represent a perturbed version of the original example $\mathrm{x_i}'$, and $d(\cdot)$ is a differentiable distance function. In this paper, we use the ***Manhattan*** distance function. Further, derive its corresponding optimal counterfactual explanation Δx^*:

$$\Delta x^* = \overset{*}{\mathrm{x_i}}{}' - \mathrm{x_i}' \tag{7}$$

The optimal counterfactual explanation Δx^* reflects the minimum change that the original example $\mathrm{x_i}'$ needs to make in order to change the model's predictions.

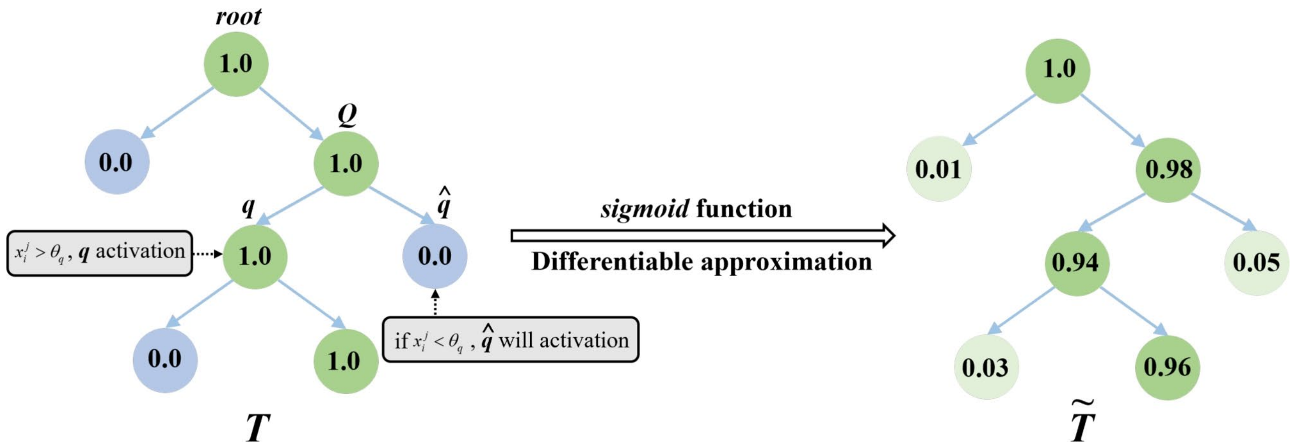
root
1.0
0.0
Q
1.0
q
$\hat{q}$
$x_i^j > \theta_q$, q activation
1.0
0.0
0.0
1.0
if $x_i^j < \theta_q$, $\hat{q}$ will activation
T
sigmoid function
Differentiable approximation
1.0
0.01
0.98
0.94
0.05
0.03
0.96
$\widetilde{T}$

Fig. 2. Simple decision tree T and its corresponding differentiable approximation tree $\widetilde{T}$.

Experiments and results

In the datasets of epilepsy (i.e. FLE vs. NC and TLE vs. NC tasks) and schizophrenia (i.e. SZ vs. NC tasks), we train tree-based classification models for brain disease diagnosis on 70% of each task: Decision Tree (DT), Random Forest (RF), and Adaptive Boosting (AB), with DT serving as the base learner. We use the remaining 30% to evaluate the model's performance and generate counterfactual examples. We have a total of 9 models (3 tasks × 3 tree-based models). It is worth noting that the hyperparameters may affect the model approximation $\widetilde{f}$, of *f*. For the selection of hyperparameters, we divide 10% of the training set as a validation set and determine the hyperparameters of the model through cross-validation. The hyperparameters we choose can produce a valid counterfactual example for each instance in the dataset and the minimum average distance between corresponding pairs $(\mathrm{x_i}', \overline{\mathrm{x_i}}')$.

We evaluate our model on two binary classification datasets: epilepsy (i.e., FLE vs. NC and TLE vs. NC tasks) and schizophrenia (i.e., SZ vs. NC tasks). We evaluate the model's performance based on diagnostic accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the curve (AUC). Among them, SEN mainly refers to the proportion of correctly predicted patients in FLE vs. NC, TLE vs. NC, and SZ vs. NC tasks, while SPE mainly refers to the proportion of correctly predicted normal controls in FLE vs. NC, TLE vs. NC, and SZ vs. NC tasks. In addition, we also measure the average distance ($dist_{\mathrm{AVG}}$) between the counterfactual examples (i.e. SC-FC coupling features after patient conversion to normal) and the original examples (i.e. SC-FC coupling features of patients) in 9 models, which reflects the average similarity between the counterfactual examples and the original examples in the feature space. A smaller $dist_{\mathrm{AVG}}$ means that the counterfactual examples are closer to the original examples in the feature space, making the model's explanation of brain diseases more intuitive and easy to understand.

As shown in Table 3, we can observe that in the three tasks, regardless of which model, the $dist_{\mathrm{AVG}}$ decreases with the increase of ACC. For example, in the FLE vs. NC task, the ACC of the RF model is better than AB, while the ACC of the DT model is better than RF, and their corresponding $dist_{\mathrm{AVG}}$ gradually decreases. In the TLE vs. NC task, the ACC of the AB model is better than RF, while the ACC of the DT model is better than AB, and their corresponding $dist_{\mathrm{AVG}}$ gradually decreases. In the SZ vs. NC task, the ACC of the AB model is better than that of the DT model, while the ACC of the DT model is better than RF, and their corresponding $dist_{\mathrm{AVG}}$ gradually decreases. This result indicates that the higher the ACC of the model, the closer the generated counterfactual examples are to the original examples. This means that a highly ACC model can better capture the SC-FC coupling features of patient anomalies, thereby generating more accurate counterfactual examples. Furthermore, we find that models with higher ACC showed significant improvements in their corresponding SEN, SPE, or AUC compared to other models.

In Figure 3, we demonstrate the disease-related abnormal brain regions captured by our model in three tasks, as well as the average perturbation level of SC-FC coupling features measured in each brain region across the three models. The average disturbance mainly represents the average degree to which the SC-FC coupling features of each brain region are altered during the generation of counterfactual explanations. If the disturbance amplitude of certain brain regions is greater than that of other regions, it indicates that these brain regions have a greater impact on the prediction results of the model. From Figure 3, it can be seen that in all three tasks, regardless of which model, the overall amplitude of the average disturbance shows the same trend. For example, in the FLE vs. NC task, the average disturbance amplitude of the ORBsup.L (Superior frontal gyrus, orbital part) and SPG (Superior parietal gyrus) brain regions is greater than that of the other regions in all three models, and the average disturbance of the ORBmid.R (Middle frontal gyrus, orbital part) region is the lowest in all three models. Similar situations are also observed in TLE vs. NC and SZ vs. NC tasks. This result further demonstrates the accuracy of the counterfactual examples generated by our model, providing strong support for our understanding of the impact of brain diseases on different brain regions. In addition, most of the brain regions we capture have shown associations with brain diseases in existing studies. For example, studies by Luo et al.[50] and O'Muircheartaigh et al.[51] indicate that FLE patients exhibit abnormalities in the left superior frontal gyrus, which corresponds to the SFGdor.L (Superior frontal gyrus, dorsolateral) and ORBsup.L regions captured in the FLE vs. NC task. Secondly, Binding et al.[52] point out that the SMA.L (Supplementary motor area) region is associated with language disorders in TLE patients, while Chabardes et al.[53] point out that TLE attacks tend to preferentially spread to the STG (Superior temporal gyrus) region. These areas are all captured in our model. This result is further validated in the schizophrenia dataset. For example, for the PCL.L (Paracentral lobule) region, Zhang et al.[54] find that SC-FC coupling in SZ patients is higher in this brain region than in other regions.

	FLE vs. NC			TLE vs. NC			SZ vs. NC		
Metric	DT	AB	RF	DT	AB	RF	DT	AB	RF
ACC(%)	84.10	81.82	82.03	85.37	82.93	80.49	82.74	84.09	80.49
SEN(%)	85.71	71.43	89.47	92.31	85.71	82.35	73.68	75.18	72.22
SPE(%)	81.82	88.24	73.91	70.59	70.26	73.68	87.50	88.24	81.25
AUC(%)	84.16	81.37	82.73	87.23	83.60	80.76	82.84	82.93	79.59
$dist_{\mathrm{AVG}}$	0.798	0.857	0.832	0.261	0.287	0.430	0.170	0.169	0.418

Table 3. Performance evaluation of 9 models. ACC: accuracy; SEN: sensitivity; SPE: specificity; AUC: area under the ROC curve; $dist_{\mathrm{AVG}}$: the average distance between the counterfactual example and the original example.

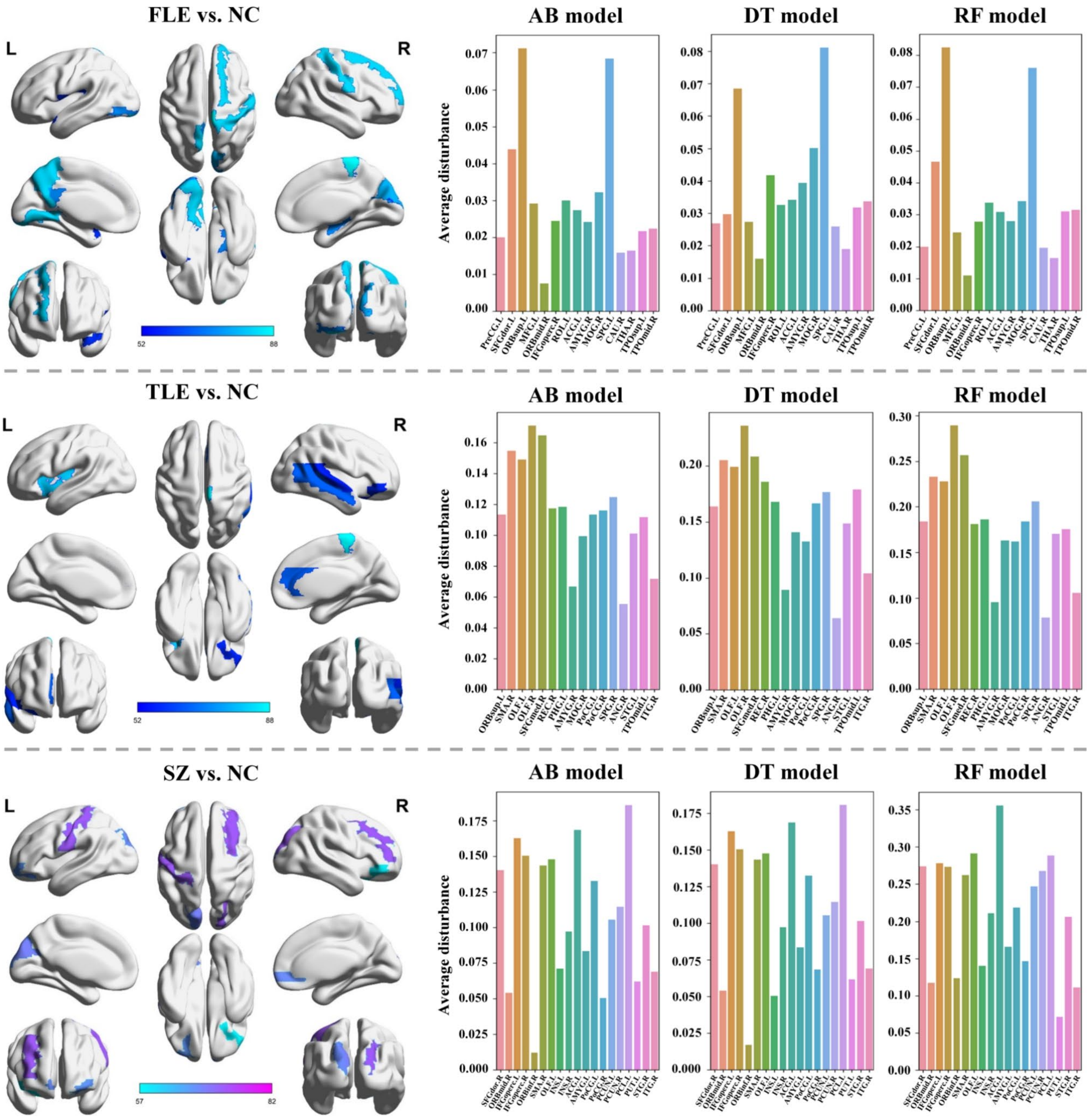

Fig. 3. The selected brain regions for three tasks and the average perturbation of each brain region corresponding to 9 models.

Besides, Ferro et al.[55] first point out abnormalities in the PoCG (Postcentral gyrus) region of SZ patients, and our model also captured this region.

To illustrate how the perturbations of different brain regions differ between different subjects, we plot a box plot as shown in Figure 4 (left), which shows the distribution of the distance (*dist*) between the counterfactual examples generated by FLE, TLE and SZ patients and the original examples in 9 models. From this figure, it can be seen that different models perform differently in generating counterfactuals in the three tasks, and *dist* fluctuates significantly between models. Specifically, in the FLE vs. NC task, the *dist* distribution of the counterfactual examples of FLE patients in the three models is concentrated between 0.6 and 0.9, which shows that the model has a strong ability to explain FLE disease and the generated results are relatively consistent. In the TLE vs. NC task, the *dist* distribution of TLE patients on the AB and DT models is tighter than that of the RF model, which shows that the RF model may have some challenges in handling the interpretation task of TLE disease, which may be due to the higher pathological complexity of TLE disease, resulting in longer *dist* of the generated counterfactual examples. In the SZ vs. NC task, the *dist* distribution on the AB and DT models is smaller than that of the RF model, which means that the RF model has a poor interpretation effect on

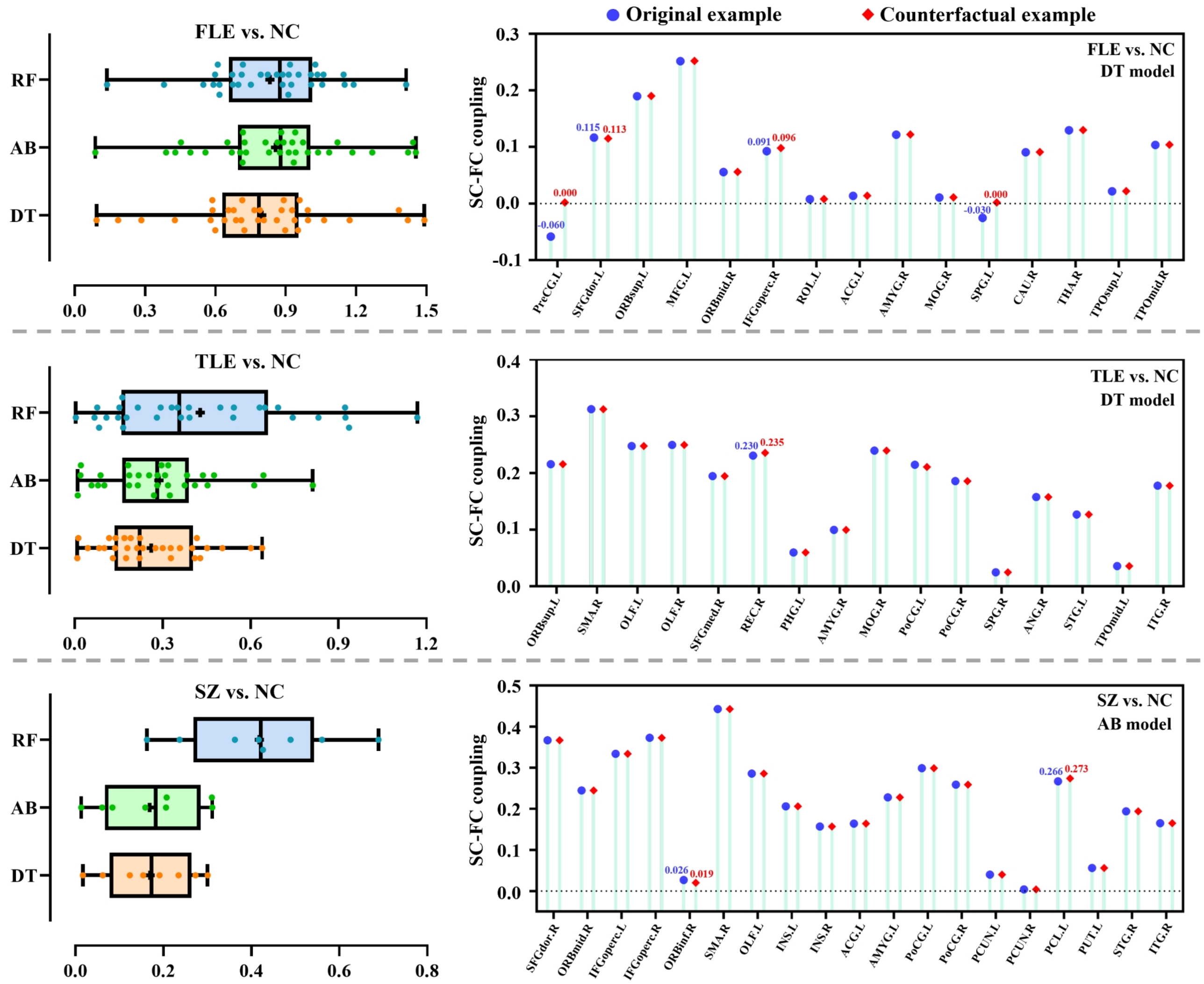

Fig. 4. Left: The distribution of the distances ($dist$) between the counterfactual examples generated by FLE, TLE, and SZ patients and the original examples in the 9 models. The black cross in the middle of each box plot indicates the average value of the distance ($dist_{\mathrm{AVG}}$). Right: Counterfactual examples generated for randomly selected FLE, TLE and SZ patients in the test set.

SZ patients. In addition, from the performance evaluation results of the models in Table 3, it can be seen that the performance of the 9 models is consistent with the above analysis. We show the counterfactual examples generated for FLE, TLE and SZ patients randomly selected from the test set in Figure 4(right). It is worth noting that this set of examples is selected based on the model with high accuracy. In the FLE vs. NC task, to bring the FLE patient to a normal state, we need to adjust its SC-FC coupling in the PreCG.L (Precental gyrus) region from −0.060 to 0.000, the SC-FC coupling in the SFGdor.L (Superior frontal gyrus, dorsolateral) region from 0.115 to 0.113, the SC-FC coupling in the IFGoperc.R (Inferior frontal gyrus, opercular part) region from 0.091 to 0.096, and the SC-FC coupling in the SPG.L (Superior parietal gyrus) region from −0.030to 0.000. In the TLE vs. NC task, we only need to adjust the SC-FC coupling of the TLE patient in the REC.R (Gyrus rectus) region from 0.230 to 0.235 to transition it to a normal state. In the SZ vs. NC task, to bring the SZ patient to a normal state, adjusting his SC-FC coupling in the ORBinf.R (Inferior frontal gyrus, orbital part) region from 0.026 to 0.019 and the ITG.R (Inferior temporal gyrus) region from 0.266 to 0.273 is necessary. The counterfactual explanation reveals the minimum changes required for input data to achieve different results[56]. It is not difficult to see that compared to the original SC-FC coupling features, the counterfactual examples we generated only make minor changes to the SC-FC coupling features. This fine-tuning, although subtle, is crucial and has great significance in the diagnosis of brain diseases.

Discussion

We explore three tree ensemble models (i.e. Decision Tree, Random Forest, and Adaptive Boosting) for the counterfactual explanation of SC-FC coupling, and to demonstrate the effectiveness of our method, we not only validate it on the epilepsy dataset, but also further empirically test it on the schizophrenia dataset. Our results show that the abnormal SC-FC coupling exhibited by patients is mainly concentrated in specific brain regions, which is consistent with the phenomenon of local abnormal connectivity in brain networks of patients with brain diseases mentioned in earlier studies. For example, the study by Chiang et al. shows that SC-FC coupling is dynamic and regulated by the duration of epilepsy[57]. Huang et al. and Liu et al. find significant differences in SC-FC coupling between TLE patients and normal subjects[58,59]. Similarly, for schizophrenia, previous studies show that the SC-FC coupling values in SZ patients are significantly higher than those in normal subjects in different brain regions[60,61]. In addition, Although previous studies have revealed the relationship between specific brain regions and their structural and functional connectivity, there is still a lack of an interpretable auxiliary diagnostic tool in the diagnosis of brain diseases. Our model can capture the brain regions where patients show abnormal SC-FC coupling and fine-tune the SC-FC coupling of the abnormal brain regions to generate counterfactual examples for the patients. To generate accurate counterfactual examples, we create an approximately differentiable version $\tilde{f}$ of the original model f, rather than replacing f with $\tilde{f}$ as in the traditional approach. This analytical method provides a useful reference for diagnosing brain diseases and offers a new perspective for understanding the nature of SC-FC coupling impairments.

However, current research still has several limitations. Firstly, we only use AAL templates to define brain regions. For future work, we will use different templates to evaluate the robustness of our proposed method[62]. Secondly, we only calculate the SC-FC coupling between each region. In future research, we will consider the impact of more factors on SC-FC coupling and conduct in-depth analysis of specific patterns or network structures in each region. Finally, for the generation of counterfactual explanations, in our experiment, only one explanatory result (one-to-one explanatory behavior) is generated for patients. We consider the possibility of generating diverse counterfactual explanations (one to many counterfactual explanations) for patients in the future.

Conclusions

In this paper, we propose a novel method for analyzing brain diseases. The existing brain disease analysis methods can only provide a simple explanation for the differences in specific brain regions between the patient group and the normal control group through statistical analysis, and cannot explain the relationship between the altered SC-FC coupling and brain diseases. Therefore, we explore three tree ensemble models (i.e. Decision Tree, Random Forest, and Adaptive Boosting) for the counterfactual explanation of SC-FC coupling. We validate the effectiveness of our approach on real epilepsy and schizophrenia datasets, and results suggest that our method can be used as a tool to assist in the diagnosis of brain diseases and provide a valuable reference for understanding the nature of SC-FC coupling impairments in these diseases.

Data availability

The data are available from the corresponding author on reasonable request.

Received: 18 June 2024; Accepted: 26 February 2025
Published online: 12 March 2025

References

1. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience* **10**, 186–198 (2009).
2. Van Den Heuvel, M. P. & Pol, H. E. H. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology* **20**, 519–534 (2010).
3. Telesford, Q. K., Simpson, S. L., Burdette, J. H., Hayasaka, S. & Laurienti, P. J. The brain as a complex system: using network science as a tool for understanding the brain. *Brain connectivity* **1**, 295–308 (2011).
4. Betzel, R. F. et al. Changes in structural and functional connectivity among resting-state networks across the human lifespan. *Neuroimage* **102**, 345–357 (2014).
5. Cociu, B. A. et al. Multimodal functional and structural brain connectivity analysis in autism: a preliminary integrated approach with eeg, fmri, and dti. *IEEE Transactions on Cognitive and Developmental Systems* **10**, 213–226 (2017).
6. Cao, W. et al. Abnormal asymmetry in benign epilepsy with unilateral and bilateral centrotemporal spikes: a combined fmri and dti study. *Epilepsy Research* **135**, 56–63 (2017).
7. Soldner, J. et al. Structural and functional neuronal connectivity in alzheimer's disease: A combined dti and fmri study. *Der Nervenarzt* **83**, 878–887 (2012).
8. Olde Dubbelink, K. T. *et al.* Disrupted brain network topology in parkinson's disease: a longitudinal magnetoencephalography study. *Brain* **137**, 197–207 (2014).
9. Liu, J. et al. Complex brain network analysis and its applications to brain disorders: a survey. *Complexity* **2017**, 8362741 (2017).
10. Li, B.-J. et al. A brain network model for depression: From symptom understanding to disease intervention. *CNS neuroscience & therapeutics* **24**, 1004–1019 (2018).
11. Fornito, A., Zalesky, A. & Breakspear, M. The connectomics of brain disorders. *Nature Reviews Neuroscience* **16**, 159–172 (2015).
12. Jie, B., Wee, C.-Y., Shen, D. & Zhang, D. Hyper-connectivity of functional networks for brain disease diagnosis. *Medical image analysis* **32**, 84–100 (2016).
13. Jie, B., Liu, M. & Shen, D. Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease. *Medical image analysis* **47**, 81–94 (2018).
14. Royer, J. et al. Epilepsy and brain network hubs. *Epilepsia* **63**, 537–550 (2022).
15. Van Diessen, E. et al. Brain network organization in focal epilepsy: a systematic review and meta-analysis. *PloS one* **9**, e114606 (2014).
16. Yasuda, C. L. et al. Aberrant topological patterns of brain structural network in temporal lobe epilepsy. *Epilepsia* **56**, 1992–2002 (2015).
17. Rubinov, M. & Bullmore, E. Schizophrenia and abnormal brain network hubs. *Dialogues in clinical neuroscience* **15**, 339–349 (2013).
18. Repovs, G., Csernansky, J. G. & Barch, D. M. Brain network connectivity in individuals with schizophrenia and their siblings. *Biological psychiatry* **69**, 967–973 (2011).
19. Aerts, H. *et al.* Modeling brain dynamics in brain tumor patients using the virtual brain. *Eneuro* **5** (2018).

20. Pan, Y. et al. Hierarchical brain structural-functional coupling associated with cognitive impairments in mild traumatic brain injury. *Cerebral Cortex* **33**, 7477–7488 (2023).
21. Liu, X. *et al.* Aberrant dynamic functional-structural connectivity coupling of large-scale brain networks in poststroke motor dysfunction. *NeuroImage: Clinical* **37**, 103332 (2023).
22. Cao, R. et al. Abnormal anatomical rich-club organization and structural-functional coupling in mild cognitive impairment and alzheimer's disease. *Frontiers in neurology* **11**, 53 (2020).
23. Piao, S. *et al.* Modular level alterations of structural-functional connectivity coupling in mild cognitive impairment patients and interactions with age effect. *Journal of Alzheimer's Disease* 1–12 (2023).
24. Popp, J. L. et al. Structural-functional brain network coupling predicts human cognitive ability. *NeuroImage* **290**, 120563 (2024).
25. Dan, T., Kim, M., Kim, W. H. & Wu, G. Uncovering structural-functional coupling alterations for neurodegenerative diseases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 87–96 (Springer, 2023).
26. Sarwar, T., Tian, Y., Yeo, B. T., Ramamohanarao, K. & Zalesky, A. Structure-function coupling in the human connectome: A machine learning approach. *NeuroImage* **226**, 117609 (2021).
27. Kulik, S. D. et al. Structure-function coupling as a correlate and potential biomarker of cognitive impairment in multiple sclerosis. *Network Neuroscience* **6**, 339–356 (2022).
28. Gu, Z., Jamison, K. W., Sabuncu, M. R. & Kuceyeski, A. Heritability and interindividual variability of regional structure-function coupling. *Nature Communications* **12**, 4894 (2021).
29. Liao, Q.-M. et al. Changes of structural functional connectivity coupling and its correlations with cognitive function in patients with major depressive disorder. *Journal of Affective Disorders* **351**, 259–267 (2024).
30. Chen, Z., Li, G., Zhou, L., Zhang, L. & Liu, J. Altered structural-functional coupling in parkinson's disease. *medRxiv* 2023–01 (2023).
31. Kong, L.-Y. et al. Divergent alterations of structural-functional connectivity couplings in first-episode and chronic schizophrenia patients. *Neuroscience* **460**, 1–12 (2021).
32. Baum, G. L. et al. Development of structure-function coupling in human brain networks during youth. *Proceedings of the National Academy of Sciences* **117**, 771–778 (2020).
33. Wei, Z., Dan, T., Ding, J., Laurienti, P. & Wu, G. Representing functional connectivity with structural detour: A new perspective to decipher structure-function coupling mechanism. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 367–377 (Springer, 2024).
34. Yang, Y. et al. Enhanced brain structure-function tethering in transmodal cortex revealed by high-frequency eigenmodes. *Nature Communications* **14**, 6744 (2023).
35. Breakspear, M. Dynamic models of large-scale brain activity. *Nature neuroscience* **20**, 340–352 (2017).
36. Bazinet, V., Hansen, J. Y. & Misic, B. Towards a biologically annotated brain connectome. *Nature reviews neuroscience* **24**, 747–760 (2023).
37. Mothilal, R. K., Sharma, A. & Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617 (2020).
38. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* **31**, 841 (2017).
39. Cheng, F., Ming, Y. & Qu, H. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics* **27**, 1438–1447 (2020).
40. Singla, S., Eslami, M., Pollack, B., Wallace, S. & Batmanghelich, K. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis* **84**, 102721 (2023).
41. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications* **11**, 3923 (2020).
42. Kim, J., Kim, M. & Ro, Y. M. Interpretation of lesional detection via counterfactual generation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 96–100 (IEEE, 2021).
43. Sirocchi, C., Urschler, M. & Pfeifer, B. Feature graphs for interpretable unsupervised tree ensembles: centrality, interaction, and application in disease subtyping. arXiv preprint arXiv:2404.17886 (2024).
44. Gangwar, M., Mishra, R. B. & Yadav, R. S. Application of decision tree method in the diagnosis of neuropsychiatric diseases. In *Asia-Pacific World Congress on Computer Science and Engineering*, 1–8 (IEEE, 2014).
45. Jin, M. & Deng, W. Predication of different stages of alzheimer's disease using neighborhood component analysis and ensemble decision tree. *Journal of neuroscience methods* **302**, 35–41 (2018).
46. Chao-Gan, Y. & Yu-Feng, Z. D. a matlab toolbox for "pipeline" data analysis of resting-state fmri. front syst neurosci. 2010; 4: 13 (2010).
47. Tzourio-Mazoyer, N. et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage* **15**, 273–289 (2002).
48. Cui, L., Bai, L., Wang, Y., Philip, S. Y. & Hancock, E. R. Fused lasso for feature selection using structural information. *Pattern Recognition* **119**, 108058 (2021).
49. Lucic, A., Oosterhuis, H., Haned, H. & de Rijke, M. Focus: Flexible optimizable counterfactual explanations for tree ensembles. *In Proceedings of the AAAI conference on artificial intelligence* **36**, 5313–5322 (2022).
50. Luo, C., An, D., Yao, D. & Gotman, J. Patient-specific connectivity pattern of epileptic network in frontal lobe epilepsy. *NeuroImage: Clinical* **4**, 668–675 (2014).
51. O'Muircheartaigh, J. & Richardson, M. P. Epilepsy and the frontal lobes. *Cortex* **48**, 144–155 (2012).
52. Binding, L. P., Dasgupta, D., Giampiccolo, D., Duncan, J. S. & Vos, S. B. Structure and function of language networks in temporal lobe epilepsy. *Epilepsia* **63**, 1025–1040 (2022).
53. Chabardès, S. et al. The temporopolar cortex plays a pivotal role in temporal lobe seizures. *Brain* **128**, 1818–1831 (2005).
54. Zhang, Z. et al. Dynamic structure-function coupling across three major psychiatric disorders. *Psychological Medicine* **54**, 1629–1640 (2024).
55. Ferro, A. et al. A cross-sectional and longitudinal structural magnetic resonance imaging study of the post-central gyrus in first-episode schizophrenia patients. *Psychiatry Research: Neuroimaging* **231**, 42–49 (2015).
56. Grath, R. M. *et al.* Interpretable credit application predictions with counterfactual explanations. arXiv preprint arXiv:1811.05245 (2018).
57. Chiang, S., Stern, J. M., Engel, J. Jr. & Haneef, Z. Structural-functional coupling changes in temporal lobe epilepsy. *Brain research* **1616**, 45–57 (2015).
58. Huang, X., Du, Y., Guo, D., Xie, F. & Zhou, C. Structural-functional coupling abnormalities in temporal lobe epilepsy. *Frontiers in Neuroscience* **17**, 1272514 (2023).
59. Liu, G. et al. Aberrant dynamic structure-function relationship of rich-club organization in treatment-naïve newly diagnosed juvenile myoclonic epilepsy. *Human brain mapping* **43**, 3633–3645 (2022).
60. Sun, Y., Dai, Z., Li, J., Collinson, S. L. & Sim, K. Modular-level alterations of structure-function coupling in schizophrenia connectome. *Human brain mapping* **38**, 2008–2025 (2017).
61. Wang, B. et al. Altered higher-order coupling between brain structure and function with embedded vector representations of connectomes in schizophrenia. *Cerebral Cortex* **33**, 5447–5456 (2023).

62. Shi, L. et al. Using large-scale statistical chinese brain template (chinese2020) in popular neuroimage analysis toolkits. *Frontiers in human neuroscience* **11**, 414 (2017).

Author contributions

S.W. and J.H. designed and conceptualized this study. Z.G. provides relevant datasets. S.W. and H.Y. conducted computational experiments, and X.Q. provided visualization of the results. S.W. and M.W. wrote and edited the manuscript, while J.H. and M.W. corrected the manuscript. All authors have reviewed and edited the original draft and approved the final submission.

Funding

This research was supported by the in part by the National Natural Science Foundation of China (No. 62471259, 62472228), the Postgraduate Research & Practice Innovation Program of School of Information Science and Technology, Nantong University (No. NTUSISTPR24_08), and the Special Research Project of Nantong Commission of Health (No. MS2023030).

Declarations

Ethical statement

The experiments on the epilepsy dataset are approved by the Ethics Committee of Nanjing University Jinling Hospital. The experiments on the schizophrenia dataset are approved by the Ethics Committee of Nanjing Medical University. All experiments are conducted in accordance with relevant guidelines and regulations. All participants provide written informed consent prior to participation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.G., M.W. or J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.