Pattern Recognition 180 (2026) 114347

Contents lists available at ScienceDirect

Pattern Recognition

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

Explainable anomaly detection and localization for ADHD in MRI using topological features☆

Peng Wang [a], Jiayi Duan [b,d], Yuqing Xing [c], Ruihang Xu [a], Anyuan Xu [a], Haodong Chen [c], Shengchao Hu [e,*], Tao Wang [b,d,*], Shuang Liu [b,d]

[a] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107, China
[b] Haihe Laboratory of Brain–Computer Interaction and Human–Machine Integration, Tianjin University, Tianjin, 300000, China
[c] School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, 430074, China
[d] Medical School, Tianjin University, Tianjin, 300072, China
[e] School of Computer Science, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Explainable AI
Attention deficit hyperactivity disorder
3D MRI
Topological data analysis
Anomaly localization

ABSTRACT

Magnetic Resonance Imaging (MRI) plays a key role in detecting physiological abnormalities. However, neurological disorders like Attention deficit hyperactivity disorder (ADHD) often lack clearly defined lesion areas, challenging traditional supervised learning approaches. Furthermore, the direct end-to-end output mechanism of traditional neural networks makes obtaining interpretable 3D features difficult, limiting clinical attribution analysis. To address these issues, we propose a topological feature-based anomaly detection framework that achieves high classification accuracy while enabling spatial recovery. Our method constructs a 3D topological space via pixel-wise pathology assessment, employs topological data analysis (TDA) to amplify subtle structural anomalies, and projects these abstract features back to the original image space using consistent positional mapping. This not only improves anomaly localization but also enhances clinical interpretability. Experiments on the ADHD-200 dataset, including site-wise and demographic subgroup evaluations, demonstrate that the proposed framework achieves strong and stable performance across multiple settings. The model also identifies lesion regions that are predominantly distributed in the prefrontal–striatal–cerebellar circuitry and further reveals biologically meaningful age- and sex-related spatial differences. This study establishes a topology-driven deep learning framework for detecting subtle anomalies and recovering their spatial distribution from 3D MRI data, offering robust data-driven support for clinical analysis.

1. Introduction

Attention deficit hyperactivity disorder (ADHD) is a prevalent neurodevelopmental disorder characterized by persistent patterns of inattention, hyperactivity, and impulsivity [1]. It affects approximately 5%–7% of children and adolescents worldwide, and its symptoms often persist into adulthood, leading to long-term impairments in academic performance, social functioning, and quality of life [2]. Despite extensive interdisciplinary research, the pathophysiological mechanisms underlying ADHD remain incompletely understood, particularly with respect to subtle abnormalities in neuroanatomical structures and brain circuits [3]. In this context, ADHD-related brain alterations can be formalized as an anomaly detection problem, where the brain representations of ADHD subjects deviate from those of healthy controls. This perspective provides a natural basis for automatic identification and localization of ADHD-related lesion regions.

Such abnormalities are often subtle, spatially distributed, and heterogeneous, making reliable detection strongly dependent on high-quality neuroimaging data. Magnetic Resonance Imaging (MRI) is widely used in neuropsychiatric research because of its non-invasive nature, high spatial resolution, and multimodal imaging capability [4]. MRI, therefore, provides an appropriate foundation for ADHD classification and lesion-region analysis [5]. However, conventional MRI-based studies often rely on predefined regions of interest, manually designed structural or functional features, or strategies based on 2D slices. These approaches are constrained by prior assumptions and expert knowledge, which limits their ability to capture complex and heterogeneous brain alterations in a unified manner.

Deep learning has recently shown promising performance in MRI-based ADHD classification by learning discriminative representations

☆ This article is part of a Special issue entitled: 'PR_Anomaly Detection, Reasoning, and Recovery' published in Pattern Recognition.

* Corresponding authors.
E-mail addresses: charles-hu@sjtu.edu.cn (S. Hu), taowang2021@tju.edu.cn (T. Wang).

https://doi.org/10.1016/j.patcog.2026.114347
Received 25 March 2026; Received in revised form 17 June 2026; Accepted 23 June 2026
Available online 27 June 2026

directly from imaging data [6]. Nevertheless, two limitations remain. First, most existing models behave as black boxes, providing subject-level predictions without explicitly identifying the lesion regions that drive the final decision [7]. This limits clinical interpretability and weakens confidence in model predictions [8]. Second, conventional neural networks mainly emphasize local spatiotemporal dependencies and are less effective at characterizing three-dimensional (3D) structural organization [9,10], such as whether abnormal responses form spatially coherent lesion regions. Such information is important for understanding the circuit-level abnormalities associated with ADHD.

These limitations motivate the introduction of Topological Data Analysis (TDA). TDA provides a principled way to characterize the global organization of high-dimensional data through connectedness and structural evolution across multiple scales [11]. Its representative technique, persistent homology (PH), captures the emergence and disappearance of topological features such as connected components during filtration, thereby yielding stable and interpretable descriptions of global data structure [12]. Owing to its robustness to noise and weak dependence on prior assumptions, TDA is particularly suitable for anomaly detection tasks that require characterization of the number, aggregation pattern, and spatial distribution of lesion regions [13]. In this paper, we utilize 3D features extracted via TDA to perform two downstream tasks: supervised classification and unsupervised detection, respectively.

Motivated by the need for both interpretability and 3D structural modeling, we propose an explainable MRI-based ADHD classification framework that integrates deep learning with topological data analysis. Specifically, a 3D MRI volume is first sliced along the sagittal, coronal, and axial directions, and a ResNet is used to classify the resulting 2D slices. The slice-level predictions from the three directions are then reorganized according to their original spatial coordinates to form a new 3D representation. On this basis, PH is applied to characterize the topological structure of abnormal responses through the Vietoris–Rips filtration and persistence barcode, and the resulting topological features are used for subject-level classification and lesion localization. Through this design, local slice-level predictions are transformed into a topological space in which PH-based features, including the number of diseased slices and the aggregation pattern of lesion regions under filtration, can be explicitly analyzed for unified subject-level classification and lesion localization.

Extensive experiments are conducted on the ADHD-200 dataset from three perspectives: full-cohort classification, site-wise evaluation, and subgroup analysis by age and sex. The proposed method achieves the best reported accuracies on five benchmark settings, namely PKU, KKI, NI, NYU, and ADHD-5, while remaining competitive on the full multi-site ADHD-200 cohort. In addition, the detected lesion regions are mainly distributed in the prefrontal–striatal–cerebellar circuitry and reveal biologically meaningful age- and sex-related spatial differences. These results indicate that the proposed framework provides both effective classification performance and interpretable lesion localization for ADHD MRI analysis. The main contributions of this work are summarized as follows:

- We propose an interpretable framework that integrates TDA with deep learning for subject-level ADHD classification and lesion region localization in 3D MRI data without relying on manually defined priors.
- We construct a unified 3D sample representation from multi-view slice predictions, and use PH-based topological features together with a volume threshold to characterize lesion-region aggregation and support subject-level diagnosis.
- We validate the proposed method on ADHD-200 through full-cohort, site-wise, and cross group experiments, and show that the identified lesion regions are consistent with known ADHD-related circuits and subgroup heterogeneity.

The remainder of this paper is organized as follows. Section 2 reviews related studies. Sections 3 and 4 introduce the preliminary concepts and the proposed methodology. Sections 5 and 6 present the experimental setup and results, respectively. Finally, Section 7 concludes the paper.

2. Related work

2.1. Attention deficit hyperactivity disorder

Attention deficit hyperactivity disorder (ADHD) is one of the most common neurodevelopmental disorders in children and adolescents. Polanczyk et al. [14] reported that the worldwide prevalence of ADHD in children and adolescents is approximately 5.29%. Li et al. [15] conducted a large-scale survey of 73,992 participants aged 6–16 in China and reported a prevalence of approximately 6.4% among Chinese children and adolescents. ADHD can impair learning ability, social functioning, and emotional well-being, and these effects often persist into adulthood. However, current diagnosis still mainly depends on clinical interviews, psychiatric manuals, and psychological testing, which are time-consuming and heavily reliant on expert judgment. Therefore, objective and efficient MRI-based diagnostic approaches are of considerable interest.

Deep learning has provided new tools for MRI-based ADHD analysis by enabling automatic feature extraction and classification from imaging data. Liu et al. [16] proposed an attention-based deep learning framework for resting-state functional Magnetic Resonance Imaging (rs-fMRI) classification. Zhang et al. [17] proposed the SC-CNN-Attention model to identify ADHD using multi-site rs-fMRI data. Aradhya et al. [18] developed a discriminative spatial filtering method to identify abnormal brain activity patterns and improve class separability. Lohani et al. [19] conducted a multimodal ADHD diagnosis study by combining features derived from structural magnetic resonance imaging (sMRI) with personal characteristics, achieving strong classification performance. Although these methods have shown promising results, most of them either rely on multimodal information or behave as black-box predictors, making it difficult to directly interpret the spatial basis of classification decisions from MRI data alone.

2.2. Topological data analysis

Topological Data Analysis (TDA) provides a mathematical framework for characterizing the global structure of data through topological invariants. Zomorodian et al. [12] established the computational foundations of PH and its use in topological analysis. Edelsbrunner et al. [20] introduced topological persistence for describing feature evolution across filtrations. Carlsson et al. [21] further developed persistence barcodes for topological characterization of high-dimensional data. These developments make HP particularly suitable for describing whether local responses remain isolated or form spatially aggregated structures.

TDA has been applied in a variety of fields, including biopharmaceuticals, finance, and image analysis. Cang et al. [13] proposed TopologyNet to integrate topological priors into image segmentation. In medical applications, Bukkuri et al. [22] reviewed the use of TDA for identifying lesion regions in medical images. Pitsik et al. [23] constructed topological disease networks to explore disease mechanisms and biomarkers. Zhang et al. [24] used persistence barcodes to characterize changes in functional-connectivity circuits in Parkinson's disease. François et al. [25] incorporated homology and representative cycles into brain MRI segmentation. However, applying TDA to MRI-based diagnosis of neurological disorders with subtle and heterogeneous abnormalities remains challenging.

Compared with the above studies, our work focuses on building an interpretable ADHD classification framework that integrates MRI-based deep learning with HP. Without relying on manually defined priors, the proposed method uses PH to characterize the aggregation pattern of lesion regions and to support both subject-level classification and lesion localization.

3. Preliminary

This section introduces the theoretical foundations of TDA, including basic definitions and notations. It also details the procedures for dataset preprocessing, as well as the construction of the discriminative space.

3.1. PH of simplicial complex

Let $S = \{s_0, \dots, s_n\}$ be the set of $n+1$ affinely independent points in the Euclidean space R^{n+1}, the *n-simplex* spanned by S is defined as a convex hull $\{\sum_{i=0}^{n} \lambda_i s_i\}$ that satisfies $\sum_{i=0}^{n} \lambda_i = 1$, $\lambda_i \geq 0$. Specifically, a single point is a 0-simplex, a line segment is a 1-simplex, a solid triangle is a 2-simplex, and a solid tetrahedron is a 3-simplex. Let K be a family of nonempty finite subsets of a point set, if K satisfies: $\tau \in K$ and $\sigma \subseteq \tau$, implies $\sigma \in K$, then we say K is an *abstract simplicial complex* [26].

For a real positive number d, the Vietoris–Rips complex is an abstract simplicial complex defined by

$$R(S,d) = \left\{ \sigma \subseteq S \;\middle|\; \text{diameter}(\sigma) \leq d \right\}, \tag{1}$$

and is referred to as the Rips complex [11]. Here, *diameter* σ denotes the maximum distance between two vertices in a simplex. Each vertex of a simplex in the Rips complex is a point, and the distance between any two vertices in the simplex does not exceed d. Given a finite point set S, we denote $d_1, \dots, d_m$ as a set of increasing non-negative real numbers, then a series of complex $\{R(S,d_1), \dots, R(S,d_m)\}$ can be constructed from S, it is called a *filtration complex* [11] if it satisfies $R(S,d_1) \subseteq \cdots \subseteq R(S,d_m)$.

Given an inclusion map $f^{i,j} : R(S,d_i) \to R(S,d_j)(j > i)$, it can induce a group homomorphism between homology groups $f_p^{i,j} : H_p(R(S,d_i)) \to H_p(R(S,d_j))$. The image of the map $f_p^{i,j}$ is called pth PH group, denoted as $H_p^{i,j}$. The pth Betti number $\beta_p^{i,j}$ is the dimension of group $H_p^{i,j}$, in other words, the rank of group $H_p^{i,j}$. Let γ be a homology class of $H_p(R(S,d_i))$, we call γ *is born in* $R(S,d_i)$ if $\gamma \notin H_p^{i-1,i}$; If γ is born in $R(S,d_i)$, we call it *dies entering* $R(S,d_j)$ if $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ and $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$. If class γ is born in $R(S,d_i)$ and dies entering $R(S,d_j)$, then γ is said to have a *persistence* of $d_j - d_i$. PH can be visually described by *persistence barcode* (PB), where the horizontal axis indicates the filtration parameter, the vertical axis signifies the Betti number of the complex at a given parameter, and the left and right ends of each bar correspond to the birthtime and deathtime of each homology class, respectively. For a dataset, the number of n-dimensional persistence bars at a fixed filtration parameter d is equal to the nth *Betti* number of the simplicial complex generated by the dataset at parameter d.

Fig. 1(a)–(e) depict the process of generating *Rips* complex based on a 2D dataset containing 7 points at filtration parameter $d = 0, 2, 4, 6, 8$, respectively. Intuitively, d also denotes the diameter of the 2D closed yellow circles. Fig. 1(f) illustrates the persistence barcode generated from these 2D points, with red and blue bars representing 0D and 1D homology, respectively. Seven cyan bars represent 7 0D homology classes and one red bar represents a 1D homology class. As the filtration parameter d increases, the number of 0D bars gradually decreases and finally only one remains; When $d = 6$, a 1D bar starts to emerge, which means a "loop" in the complex is generated by the point set under the filtration parameter. However, as d further increases, the "loop" soon disappears.

3.2. Data splitting and local discrimination

For a 3D MRI dataset with sample size N, we denote it as $\{(X,y)_i\}_{i=1}^{N}$, where $X \in R^{m\times n\times t}$ is the feature, $m,n,t \in \mathcal{N}^+$ represents the number of slices along sagittal, coronal, and transverse plane in X respectively, and $y \in \{0,1\}$ is the corresponding label. For convenience, let X_k^j denote the jth position after the cuts along the kth cutting section($j \in \mathcal{N}^+, k = 1,2,3$). Besides, let y_k^j be the label of X_k^j; it accords with the label of the X, i.e., all slices from the same MRI volume share the same label, which corresponds to a weakly supervised setting. According to the above slicing, we carry out the same process for all the samples in the MRI dataset. Thus, we obtain subdatasets $\{(X_k^j, y_k^j)_i\}_{i=1}^{N}$ for each position with size N, which consist of the slices from the same position. Each slice in the above datasets corresponds to a 2D image, for instance, $X_1^j \in R^{n\times t}$.

Based on the above 2D grayscale image datasets, we train the ResNet for classification, and the output is notated as $\hat{y}_k^j \in \{0,1\}$. We join together all the classification results on the kth cutting section in order of position, denoted by the discriminant vector $\hat{y}_k = (\hat{y}_k^1, \hat{y}_k^2, \dots, \hat{y}_k^{j_k})$, where $j_1 = m$, $j_2 = n$, and $j_3 = t$. Thus, we combine three discriminant vectors and represent them in a 3D Cartesian coordinate system, referring to the resulting structure as the discriminant tensor $\hat{Y} \in R^{m\times n\times t}$ of X. And the pixel value at any spatial position and direction within the original feature X can be encoded as a binary value $\{0,1\}$ that contains decision information.

3.3. Discriminant space

Next, we analyze the ADHD data for PH characteristics. We define the 1D discriminant space ${}^1\mathrm{P}_k \subseteq R$ to be the set of all 1D points j satisfying $\hat{y}_k^j = 1$ at the kth cutting section in an MRI sample. Additionally, we introduce $L_j(d)$ to represent the line segment centered at site $j \in {}^1\mathrm{P}_k$ with length equal to $d(d \geq 0)$. Then, the union of these line segments is denoted as

$$U_{{}^1\mathrm{P}_k}(d) = \bigcup_{j\in {}^1\mathrm{P}_k} L_j(d). \tag{2}$$

In fact, $U_{{}^1\mathrm{P}_k}(d)$ is a topological space under the Euclidean topology, and we can divide it into several connected components $\{U_1, \dots, U_m\}, m \in \mathcal{N}^+$. Considering practical scenarios, when $d < 1$, the number of connected components in $U_{{}^1\mathrm{P}_k}(d)$ is the number of points in ${}^1\mathrm{P}_k$, which can represent the number of slices judged to be diseased in the kth cutting section of a sample. Based on the above, we define the 3D discriminant space as the product space

$${}^3\mathrm{P} = {}^1\mathrm{P}_1 \times {}^1\mathrm{P}_2 \times {}^1\mathrm{P}_3. \tag{3}$$

As we can see, the 3D discriminant space ${}^3\mathrm{P}$ is a set of 3D points satisfying $(\hat{y}_1^j, \hat{y}_2^j, \hat{y}_3^j) = (1,1,1)$, which denotes the component values of the jth position in three directions are all 1. In practical applications, each point in ${}^3\mathrm{P}$ describes an intersection position of three diseased slices of different cutting sections in the MRI data. In other words, this location is identified as diseased in all three directions. Subsequently, we compute the topological persistence of discrete points in the discriminant space ${}^3\mathrm{P}$ by constructing the *Rips* complex and obtaining persistence barcodes of these points.

4. Methodology

This model comprises two main components: (1) local discrimination achieved through 3D slicing and the generation of a new topological space, and (2) the application of TDA within this space, enabling localization and classification based on extracted topological features, as illustrated in Fig. 2. In the first part, we slice the 3D MRI images into 2D images along the sagittal, coronal, and transverse planes. Then, we classify them using ResNet [27] to construct a stable and unambiguous discriminant, which preserves the original spatial information and facilitates the extraction of topological features in subsequent analyses. Following that, PH is leveraged to analyze and obtain interpretable topological features that can differentiate between normal and diseased samples. Finally, we introduce an algorithm based on topological features for classification tasks and use it to detect lesions. In the following subsection, we will describe each part in detail.

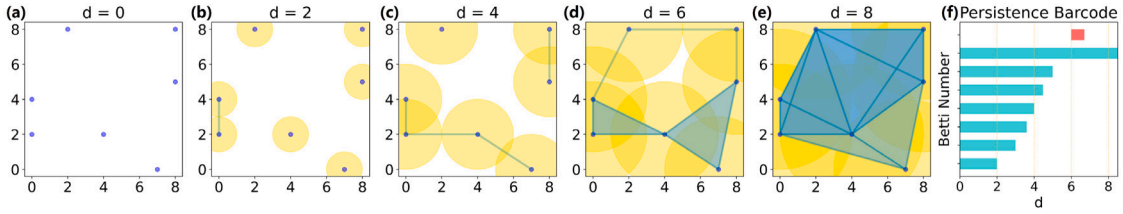

Fig. 1. (a)–(e) show the *Rips* complex generated by a 2D dataset containing 7 points at filtration parameter $d = 0, 2, 4, 6, 8$ respectively; the complex at $d = 8$ has become a full complex, by continuing to increase the filtration parameter, the homology of the generated complex is no longer changing. (f) shows the persistence barcode generated by these 7 points, with red and cyan bars representing 0D and 1D homology, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

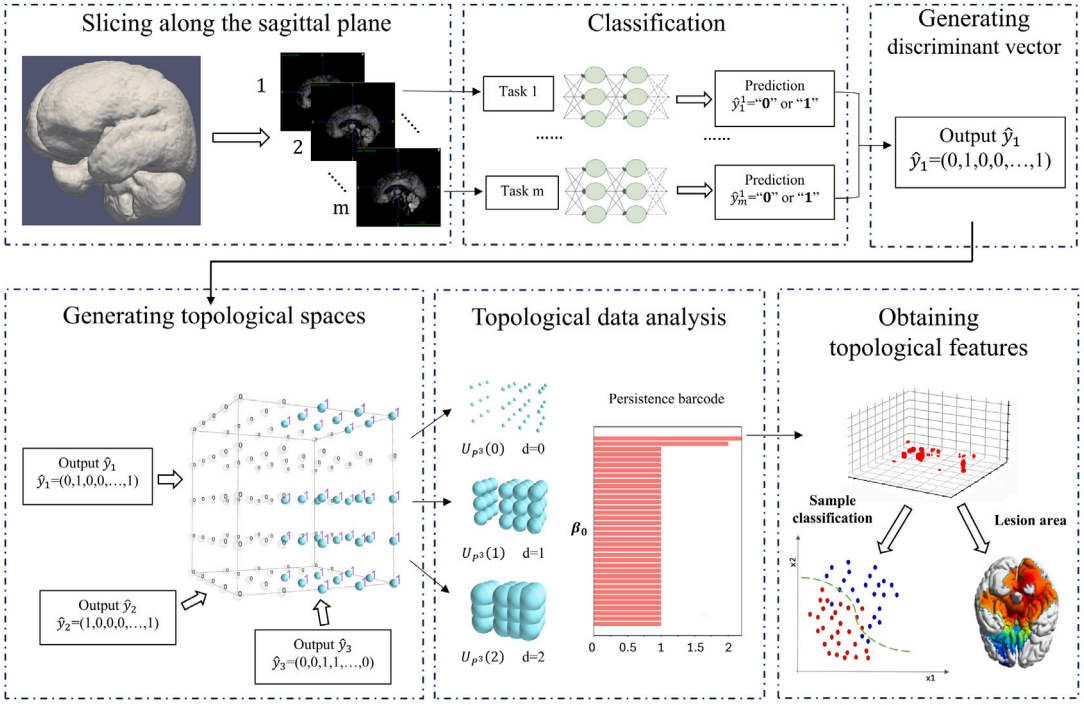

Fig. 2. The overall pipeline of the model, including two steps. For an input 3D MRI data, firstly it is sliced and locally discriminated by the ResNet to form a new data space ${}^3\mathrm{P}$, $U_{{}^3\mathrm{P}}(d)$ represent the union of closed spheres with radius $d/2$ centered on all points in ${}^3\mathrm{P}_k$. Then TDA is employed to capture topological characteristics within $U_{{}^3\mathrm{P}}(d)$, finally the algorithm based on topological features outputs classification results and is capable of recognizing lesions.

4.1. Topological data analysis

In the previous section, we transform 2D gray images of X into binary signals with discriminative information for local determination. This section will look for topological features to identify positive and negative samples in discriminant spaces utilizing topological data analysis (TDA), where PH is mainly used. PH studies the change of topological features in the filtration process by building complexes with different parameters of discrete datasets. It can be depicted by the persistence barcode [28], from which we can obtain the duration of topological features as parameters continuously change. Subsequently, these identified features are employed for subsequent downstream tasks.

4.1.1. Topological features

We have identified two topological features based on PH: the number of diseased slices and the distance between diseased slices in MRI data. In light of this, we describe the difference between the two features using the persistence barcode of a diseased sample and a normal sample in Fig. 3. It shows the 0D PH features of diseased and normal samples in the sagittal 1D discriminant space and 3D discriminant space, respectively, where red and orange horizontal bars describe the 0D PH of the diseased and normal samples, respectively.

In the picture, the diseased sample generates significantly more bars than the normal sample. Specifically, in Fig. 3(a), when $d = 0$, the number of red and orange bars is 13 and 3, respectively, indicating that 13 and 3 slices in the sagittal section of the diseased and normal samples are identified to be diseased respectively; When $d = 1$, although the three red bars disappear, orange bars continue to exist, and the difference of number between the two remains significant. This phenomenon becomes even more pronounced in 3D space. In Fig. 3(b), at $d = 0$, the top red bar has a y-coordinate of 4576, while the top orange bar's y-coordinate is 45. Due to the limitation of space, we collapse part of the information on the y-axis. When transitioning to $d = 1$, although 2886 red bars vanish, a notable disparity persists in the number of remaining bars between the normal and diseased samples, specifically 1690 versus 45.

Besides, when d changes continuously, there is a significant change in 0D homology classes of diseased samples, whereas that of normal samples remains insensitive to the change. This indicates that it is easier to form blocks of close points in the discriminant space in the diseased sample while the normal sample exhibits discrete points. More specifically, the disappearance of the three red bars in Fig. 3(a) as d

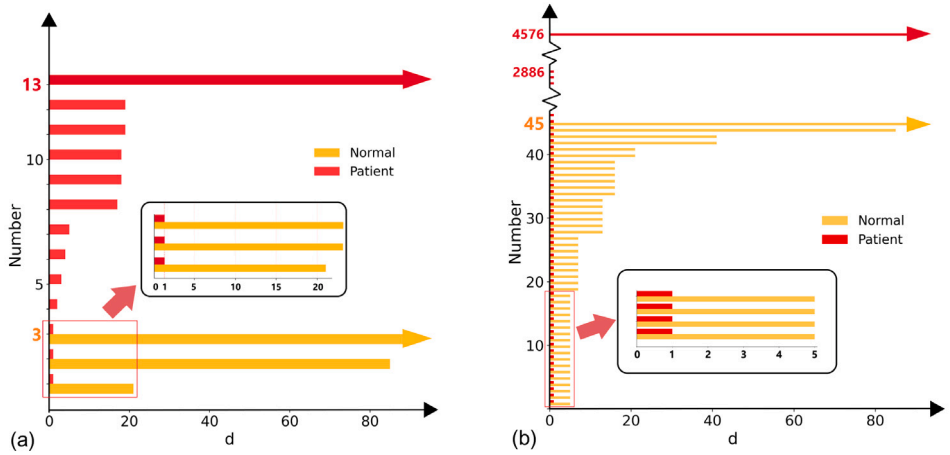

Fig. 3. (a) shows topological distinctions between a normal sample and a diseased sample in the sagittal (space ${}^1\mathrm{P}_1$) . (b) depicts the overall topological differences in the 3D discriminant space ${}^3\mathrm{P}$ between them. The horizontal coordinate represents the filtration parameter d and the vertical coordinate represents the number of 0D bars generated in the filtration process. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

goes from 0 to 1 implies that there are slices that are judged to be diseased at all neighboring locations and that these slices can form the area considered to be diseased. In contrast, the first orange bar disappears until $d = 21$, which indicates that the minimum distance between diseased slices in the sample is 21, which prevents them from forming connected regions. Fig. 3(b) shows the same phenomenon in 3D space. Let d_{min} represent the smallest distance between the intersection locations of diseased slices that are judged as diseased in all three directions within the sample. Although d_{min} in the normal sample decreases to 5, 2886 red bars in the diseased sample disappear when $d = 1$, implying that larger connected regions appear in the diseased sample.

In summary, the above two topological features describe the possibility of a sample being diseased and the existence of connected diseased regions, respectively, and they can basically distinguish diseased samples from normal ones. We further consider the size of a connected region to achieve a more precise distinction. For example, when $d = 1$, the homology classes represented by "11" and "111" (where a "1" indicates a position identified as diseased) are equivalent. However, the connected regions that they form exhibit differences in size. Therefore, we introduce C as an additional feature to represent the number of adjacent slices identified as diseased. If the distance between two points in the discriminant space is equal to 1, we call the two points adjacent. Based on this, we set a classification criterion in a single direction to show the role of C: for the kth section of a sample, if the length $L(U_m)$ of the connected component U_m in the topological space $U_{1\mathrm{P}_k}(d)$ satisfies $L(U_m) \geq C$, the sample is classified as diseased; otherwise, it is classified as normal.

4.2. Classification algorithm based on topological features

To construct algorithms based on the above topological features to complete the final classification, we further focus on the change in the generated 3D topological space as parameter d changes from 0 to 1. This is because we have experimentally verified that $d = 1$ is the optimal parameter for a single direction (see Table 5). Let $U_{3\mathrm{P}}(d)$ represent the union of closed spheres with radius $\frac{d}{2}$ centered on all points in ${}^3\mathrm{P}$, i.e., $U_{3\mathrm{P}}(d) = \bigcup_{x\in{}^3\mathrm{P}} B(x, \frac{d}{2})$, where $B(x, \frac{d}{2}) = \{x' \in {}^3\mathrm{P} \mid \mathrm{distance}(x', x) \leq d\}$.

When $0 \leq d < 1$, the topological space formed by all points consists of a union of isolated solid spheres; When $d = 1$, spheres clinging closely with each other in ${}^3\mathrm{P}$ lie in a common connected component topologically. In other words, all adjacent points are contained in this connected region. This corresponds to whether there is an observable blocky lesion in the sample, which is also the key point for our determination of whether the sample is normal. Therefore, this setting captures local spatial continuity while avoiding the connection of distant and potentially unrelated positive responses. This provides a natural criterion for identifying contiguous lesion-region candidates in the reconstructed 3D space.

As shown in Fig. 3(b), space $U_{3\mathrm{P}}(1)$ may contain multiple connected components $U_m, m \in \mathcal{N}^+$. We target the connected components with the largest size and set the threshold $V_0 = \frac{\pi \prod_{k=1}^{3} C_k}{6}$, where $k = 1, 2, 3$, and C_k is the threshold in the kth direction(as the C we set before). We use this to set the criterion for determining: if there exists connected component U_m in the topological space $U_{3\mathrm{P}}(1)$ satisfying $V(U_m) \geq V_0$, the sample is classified as diseased; otherwise, it is classified as normal. Thus, V_0 represents the minimum spatial extent required for a connected component to be regarded as a lesion region. This threshold prevents small isolated clusters from directly determining the subject-level diagnosis. Finally, we provide the classification algorithm 1.

Algorithm 1 Classification algorithm based on topological features

Input: $U_{3\mathrm{P}}(d)$ generated by discriminant tensor $\hat{Y}$;
Output: Sample classification result $\hat{y}$;

```
Initialize: d = 1, C_k = 2(k = 1, 2, 3)
for U_m ∈ U_3P(1), m = 1, 2, ... do
    if V(U_m) ≥ π∏_{k=1}^{3} C_k / 6 then
        ŷ = 1
        break
    else
        ŷ = 0
    end if
end for
return ŷ
```

In the algorithm, U_m represents the mth connected component of $U_{3\mathrm{P}}(d)$, and $V(U_m)$ denotes the volume of the geometric realization of U_m in the 3D Euclidean space $\mathbb{R}^3$. In fact, algorithm 1 classifies the samples by determining whether there are lesion regions of larger size in the space, where the judgment is based on volume threshold V_0. In addition, we consider C_k as the hyperparameter to control the size of the threshold V_0, i.e., the size of the lesion region, by tuning C_k.

Table 1
Summary of the ADHD-200 dataset used in this study, including the overall cohort, site-wise subsets, and demographic subgroup partitions. The ADHD-5 dataset consists of the PKU, KKI, NI, NYU, and OHSU sub-datasets.

Category	Subset	ADHD	TD	Total
Site	PKU	101	143	244
	KKI	25	69	94
	NI	36	37	73
	NYU	151	108	259
	OHSU	42	69	111
	UPit	4	94	98
	WashU	0	59	59
Age	Child	215	329	544
	Adolescent	144	250	394
Gender	Female	77	277	354
	Male	281	302	583
Overall	ADHD-200	359	579	938

5. Experiment

5.1. Datasets

We evaluate the proposed framework on the ADHD-200 dataset from three perspectives: full-cohort classification, site-wise evaluation on PKU, KKI, NI, NYU, and ADHD-5, and subgroup analysis by age and gender. The ADHD-200 dataset consists of sub-datasets from eight different research institutions, including 579 controls (TD, typically developing) and 359 ADHD patients, which are delineated and available at http://fcon_1000.projects.nitrc.org/indi/adhd200/, together with phenotypic annotations such as diagnosis, age, and gender [29]. The ADHD-5 dataset consists of the PKU, KKI, NI, NYU, and OHSU sub-datasets. The task is a binary classification of healthy controls and ADHD patients, and the size of each MRI sample is $121 \times 145 \times 121$. The detailed data distribution is summarized in Table 1.

To further investigate demographic heterogeneity, we additionally conduct subgroup analyses with respect to gender and age. After excluding one subject with missing gender annotation, the gender-stratified experiments are conducted on 937 subjects, including 354 females (77 ADHD and 277 TD) and 583 males (281 ADHD and 302 TD). For age-stratified experiments, we follow the clinical relevance of early-onset ADHD and define a Child group covering ages 6–12 years [30] and merge all subjects older than 12 years into a single older group, denoted as Adolescent in the following experiments. This results in 544 Child subjects (215 ADHD and 329 TD) and 394 Adolescent subjects (144 ADHD and 250 TD), as also summarized in Table 1.

5.2. Implementation details

For 2D slice classification, we use an ImageNet-pretrained ResNet34 as the backbone to train on the 2D image dataset $\{(X_k^j, y_k^j)_i\}_{i=1}^N$. Before slice-level training, uninformative boundary regions are removed to reduce background interference. Each 2D slice is center-cropped to 120×120, and intensity normalization is applied to reduce inter-subject intensity variation. These preprocessing steps ensure that the input slices have consistent spatial dimensions and comparable intensity distributions before being fed into the ResNet classifier. As summarized in Table 2, the model is trained using the Adam optimizer and cross-entropy loss, with a learning rate of 4×10^{-3} and a weight decay of 1×10^{-8}. Early stopping with a patience of 20 epochs is adopted to alleviate overfitting.

For 3D sample-level classification, we optimize the topological parameters C_k together with the filtration parameter d to determine the volume threshold V_0 for lesion-region detection. In all experiments, the best performance is achieved with $d = 1$, while the optimal choices of C_k are reported in Section 6.3. Lesion regions identified by the model are visualized using BrainNet Viewer [31]. Accuracy, sensitivity, and specificity are used as the evaluation metrics. The entire framework is assessed under a stratified five-fold cross-validation scheme, with final performance reported as the average across the five folds.

Table 2
Summary of hyperparameters.

Parameter	Value	Parameter	Value
Loss function	Cross Entropy	Early Stop Steps	20
Optimizer	Adam	Weight Decay	1e−8
Learning rate	4e−3	Dropout	0.1
Batch size	64	Epochs	50

All experiments were implemented on a single NVIDIA RTX 4090 GPU (40 GB). For the full ADHD-200 cohort, the training stage required approximately 5–15 GPU-hours in total, with a peak device memory of 4–8 GB; this offline cost is amortized over the cohort. In the test stage, end-to-end processing of one subject required approximately 0.3–1.2 s per subject, of which ResNet34 forward passes accounted for approximately 0.2–0.8 s and Algorithm 1 for less than 0.1 s, with peak memory below 2 GB. Scoring the full cohort of 938 subjects required approximately 5–20 min on GPU; per-fold storage of discriminant outputs was under 1 MB.[1]

6. Results

In this section, we report the experimental results from three perspectives: comparison with existing methods, the effect of topological parameters, and explainability analysis of lesion-region localization.

6.1. Comparison with different methods

Table 3 compares the proposed method with representative previous studies on the ADHD-200 dataset and its commonly used subsets. Accuracy is adopted as the primary metric for comparison, and dashes indicate that the corresponding result was not reported. It should be noted that existing methods may differ in preprocessing strategies, input modalities, data splits, and validation protocols. Therefore, the results in Table 3 are intended to demonstrate the competitiveness of the proposed framework rather than to claim absolute superiority under identical experimental conditions.

Overall, the proposed framework shows strong competitiveness across different evaluation settings, particularly on the independent site-specific subsets, namely PKU, KKI, NI, NYU, and ADHD-5. Although it does not surpass the best previously reported result on the full ADHD-200 benchmark, it remains competitive under substantially stronger inter-site heterogeneity. Compared with single-site subsets, the complete ADHD-200 cohort involves broader variations in acquisition protocols, scanner characteristics, demographic composition, and clinical distributions, making classification considerably more challenging. Therefore, the site-wise experiments are important for evaluating whether the proposed framework can maintain stable performance under different data distributions, which will be reported in Section 6.3.

Another notable advantage is that most previous studies reported results on only a limited number of subsets, whereas the proposed framework provides a unified evaluation across all six settings in Table 3. This suggests better adaptability to datasets with different sample sizes and distribution characteristics. In particular, the strong performance on relatively small subsets indicates favorable data efficiency, which is important for neuroimaging studies where large-scale labeled MRI data are difficult to obtain.

[1] Interested readers can contact the authors for access to the code, and we will be happy to provide the necessary resources.

Table 3
Classification accuracy (%) of different methods on the ADHD-200 dataset and its five sub-datasets. Our work is compared with previous state-of-the-art methods using the same datasets. The dash '–' in the table indicates that the corresponding result was not reported. Bold indicates the best performance for each dataset.

Research	Method	PKU	KKI	NI	NYU	ADHD-5	ADHD-200
Farzi et al. [32]	DBN	–	–	69.83	63.68	–	–
Zou et al. [33]	3DCNN	62.95	72.82	–	70.50	–	69.15
Riaz et al. [34]	DeepFMRI	62.70	–	67.90	73.10	–	–
Zhang et al. [17]	SC-CNN	65.20	77.70	75.30	60.40	68.60	–
Aradhya et al. [18]	DSFM	–	–	–	–	–	73.83
Khan et al. [35]	KD-FS	60.00	72.00	70.00	73.00	–	–
Wang et al. [36]	Slice-RF	–	81.82	–	70.73	–	**75.46**
Liu et al. [16]	ConvGRU	–	–	–	–	72.44	–
Li et al. [37]	BNPTT	–	–	–	–	–	71.14
Ours	ResNet-TDA	**66.67**	**83.33**	**76.18**	**75.61**	**74.07**	74.53

Table 4
Performance comparison between majority voting baselines and the proposed ResNet-TDA.

ResNet-TDA (Ours)	MVC, $X = 0$	$X = 10$	$X = 15$	$X = 20$	$X = 25$
74.53	37.97	54.01	65.24	71.66	71.66

These results also highlight the contribution of the PH-based subject-level decision stage beyond slice-level classification. By reorganizing multi-directional slice predictions into a new sample space and analyzing their topological evolution under the Vietoris–Rips filtration, the proposed method suppresses isolated noisy slice responses and improves the reliability of subject-level diagnosis. This property is particularly beneficial for relatively small or homogeneous subsets, while still maintaining competitive performance on the more challenging multi-site ADHD-200 benchmark.

6.2. Ablation study for the TDA decision module

To thoroughly evaluate the necessity and effectiveness of the proposed TDA module, we conduct an ablation study comparing it with a majority voting classifier (MVC). The MVC baseline aggregates slice-level predictions from the three orthogonal planes and determines the subject-level diagnosis based on the proportion of positively predicted slices, defined as

$$\hat{y} = \begin{cases} 1 \text{ (ADHD)}, & \text{if } r \geq X\% \\ 0 \text{ (Control)}, & \text{otherwise}, \end{cases} \tag{4}$$

where $r = \frac{\sum_{j=1}^{121} y_1^j + \sum_{j=1}^{145} y_2^j + \sum_{j=1}^{121} y_3^j}{121+145+121}$ denotes the proportion of positively classified slices after aggregating predictions from the three planes. A subject is classified as ADHD if the proportion r exceeds the predefined threshold $X\%$.

Table 4 summarizes the classification performance of the proposed ResNet-TDA framework and the MVC baselines under different threshold settings. As shown in Table 4, the MVC achieves its best performance at $X = 20\%$, reaching an accuracy of 71.66%. Nevertheless, the proposed ResNet-TDA consistently outperforms all MVC configurations, achieving a superior accuracy of 74.53%.

The performance gain primarily arises from the fundamental difference in the decision mechanisms of the two approaches. The MVC baseline relies solely on the scalar ratio r of positively predicted slices, resulting in a one-dimensional threshold-based decision rule. Consequently, it is sensitive to spurious consistency, where isolated false-positive slices may accumulate numerically despite lacking meaningful spatial continuity. In contrast, the proposed TDA module explicitly models the three-dimensional adjacency relationships within the decision tensor by capturing spatially connected topological components. This enables ResNet-TDA to suppress scattered, non-cohesive false positives while preserving physically contiguous abnormal regions. Therefore, replacing a simple slice-ratio criterion with a topology-aware cluster evaluation substantially improves robustness and interpretability at the subject level.

Table 5
Classification accuracy (%) of different C under different filtration parameters along the sagittal direction.

	C = 1	C = 2	C = 3	C = 4	C = 5
d = 1	45.03	52.63	65.50	**71.35**	70.76
d = 2	45.03	49.71	54.97	64.91	70.18
d = 4	45.03	60.23	60.23	60.23	60.23

6.3. Effect of topological parameters

This subsection analyzes the influence of the topological parameters on classification performance, with particular emphasis on the threshold parameter C (or C_k in 3D) and the filtration parameter d. We first consider a simplified 1D setting along the sagittal direction to illustrate how different combinations of C and d affect the PH-based decision process. We then extend the analysis to the topological space constructed from the new sample space and report the optimal parameter configurations for both site-wise and subgroup-level evaluations.

Table 5 reports the average classification accuracy under five-fold cross-validation for different values of C and d along the sagittal direction. At one extreme, when $C = 1$, a sample is predicted as diseased once at least one slice is predicted as positive. In this case, all samples are assigned to the positive class, and the accuracy remains 45.03% regardless of d. When $d = 1$, increasing C imposes a stricter continuity requirement on positive slices. The best performance is obtained at $C = 4$, corresponding to the condition $L(U_m) \geq 4$, where the connected component must span at least four consecutive positive slices. In contrast, when $d = 4$, the spatial constraint becomes more relaxed and the accuracy stabilizes at 60.23%. These results indicate that the choice of C and d directly controls the trade-off between sensitivity to local lesion clusters and robustness against isolated false positives.

We next tune the parameters C_k in the topological space derived from the new sample space. Table 6 summarizes the optimal configurations of C_1, C_2, and C_3 for each dataset, together with the corresponding mean classification results. Different datasets favor different topological thresholds: relatively larger configurations are selected for NI and NYU, whereas more compact settings are preferred for KKI and ADHD-5. This indicates that the proposed PH-based topological analysis can adapt to different levels of heterogeneity and lesion-region aggregation patterns.

As shown in Table 6, the optimized parameters yield competitive and generally well-balanced performance across all datasets. In particular, sensitivity and specificity remain closely matched in most settings, suggesting that the proposed criterion does not strongly bias the classifier toward either class. This trend is also reflected in the F1-scores reported in Table 6, which confirm that the proposed framework maintains a balanced trade-off between precision and recall for the ADHD class across different datasets. This observation supports the role of the topological decision stage in improving subject-level robustness beyond slice-level predictions.

Table 6
Experimental settings and classification performance for the ADHD classification task. C_1, C_2, and C_3 denote the thresholds applied to the sagittal, coronal, and axial directions, respectively. For each dataset, we report the optimal topological thresholds and the corresponding mean classification results.

	PKU	KKI	NI	NYU	ADHD-5	ADHD-200
$C_1 \times C_2 \times C_3$	$5 \times 6 \times 2$	$2 \times 2 \times 3$	$5 \times 5 \times 5$	$6 \times 4 \times 6$	$2 \times 2 \times 3$	$3 \times 3 \times 3$
Accuracy (%)	66.67	83.33	76.18	75.61	74.07	74.53
Sensitivity (%)	67.33	84.00	76.20	76.16	74.08	74.60
Specificity (%)	66.43	82.61	76.16	75.00	73.94	74.49
F1-score (%)	62.67	72.41	75.93	78.50	72.15	69.16

Table 7
Experimental settings and classification performance for the subgroup analysis of the ADHD classification task. C_1, C_2, and C_3 denote the thresholds applied to the sagittal, coronal, and axial directions, respectively. For each subgroup, we report the optimal topological thresholds and the corresponding mean classification results.

	Child	Adolescent	Female	Male
$C_1 \times C_2 \times C_3$	$4 \times 3 \times 4$	$2 \times 3 \times 4$	$3 \times 3 \times 3$	$5 \times 3 \times 3$
Accuracy (%)	73.47	64.62	78.69	64.38
Sensitivity (%)	73.35	64.58	78.65	64.41
Specificity (%)	73.55	64.80	78.70	64.24
F1-score (%)	68.61	57.23	61.62	63.51

We further apply the same analysis to demographic subgroups. Table 7 reports the optimal configurations of C_1, C_2, and C_3 for the Child, Adolescent, Female, and Male subgroups, together with the corresponding mean classification results. Again, different subgroups favor different threshold combinations, indicating that the spatial aggregation pattern of lesion regions varies across demographic partitions.

As shown in Table 7, the subgroup results are also generally balanced in terms of sensitivity and specificity. The Female and Child subgroups achieve relatively stronger performance, whereas the Adolescent and Male subgroups are more challenging. Nevertheless, the comparable sensitivity and specificity values across subgroups indicate that the proposed topological criterion preserves balanced discrimination under different demographic settings. This trend is further supported by the F1-scores, which remain relatively stable across subgroups and confirm that the proposed framework maintains a balanced trade-off between precision and recall for the ADHD class.

Overall, the results from the sagittal analysis to the full 3D setting show that the topological parameters play a key role in balancing local sensitivity and global robustness. Small thresholds tend to overemphasize isolated positive slices, whereas overly large thresholds may suppress genuinely discriminative local structures. By selecting dataset-specific combinations of C_1, C_2, and C_3, the proposed framework achieves stable and balanced classification performance across multiple datasets.

6.4. Computational complexity analysis

On an NVIDIA RTX 4090 GPU, during the training phase, completing five-fold cross-validation and generating decision tensors requires a total of 5–15 h, with a peak GPU memory footprint of 4–8 GB. This offline computational overhead is amortized over the entire cohort. During the inference (testing) phase, processing a single subject requires only 0.3–1.2 s (comprising 0.2–0.8 s for the ResNet34 forward pass and less than 0.1 s for the decision-making process in Algorithm 1), with a peak memory consumption of less than 2 GB. Scoring and inference for the entire cohort of 938 subjects takes merely 5–20 min, requiring a storage footprint of less than 1 MB per fold. These computational specifications fully satisfy the practical deployment requirements for offline clinical screening assistance.

The primary computational cost during training consists of discriminative tensor generation (387 ResNet34 forward passes per subject, taking approximately 0.8 s on an RTX 4090) and five-fold slice-level training (roughly 12 GPU-hours on the ADHD-200 cohort, $N = 938$). Tuning the topological threshold on precomputed predictions $\hat{Y}$ requires only nearly 3.5 min for the complete five-fold grid search (125 candidate configurations of (C_1, C_2, C_3) per fold across around 150 validation subjects), representing approximately 0.5% of the training runtime. At inference, Algorithm 1 introduces minimal overhead (about 4 ms per subject) compared to the slice-level CNN passes (roughly 800 ms), constituting merely 0.5% of the total per-subject latency.

6.5. Explainability analysis

This subsection investigates the interpretability of the proposed framework, with the aim of clarifying the decision basis for ADHD identification and exploring the neurobiological significance of the lesion regions detected by the model. The analysis is conducted from three perspectives: lesion-region distributions in different sub-datasets, probability distributions in the full ADHD-200 cohort, and age- and sex-related subgroup comparisons. Together, these analyses provide further insight into the PH-based decision mechanism and the spatial distribution patterns of ADHD-related lesion regions.

6.5.1. Analysis of lesion regions in sub-datasets

We first examine four representative sub-datasets of ADHD-200, namely PKU, KKI, NI, and ADHD-5, to determine whether the lesion regions identified by the model exhibit subset-specific spatial patterns. For each subset, we construct a subgroup-level lesion probability map and retain lesion regions with occurrence probability greater than 50%. The corresponding visualizations are shown in Fig. 4.

As shown in Fig. 4(a)–(c), the lesion regions identified in PKU, KKI, and NI are mainly distributed in the frontal lobe, particularly around the dorsolateral prefrontal cortex, ventromedial prefrontal cortex, and anterior cingulate cortex. These regions are closely related to executive control, emotion regulation, and conflict monitoring [38]. Their involvement is consistent with previous studies reporting frontal structural and functional abnormalities in ADHD [39].

In addition, Fig. 4(b)–(d) shows that frontal lesion regions are predominantly concentrated in the left prefrontal cortex, suggesting a degree of hemispheric asymmetry. This observation is in line with previous findings of altered lateralization in ADHD [40]. As shown in Fig. 4(b) and (d), lesion regions in ADHD-5 are also prominently distributed in the striatum and cerebellum. This finding agrees with prior evidence highlighting the involvement of these regions in ADHD [41]. Abnormalities in these regions may underlie reward-processing deficits, attentional instability, and impulsive behavior in ADHD [42].

6.5.2. Probability distribution of lesion regions in the ADHD-200 dataset

To further investigate the population-level characteristics of lesion regions, we analyze their probability distribution across the entire ADHD-200 dataset. Specifically, we compute the occurrence probability of each detected lesion coordinate across all ADHD subjects and visualize the resulting probability map in Fig. 5.

As shown in Fig. 5, the dominant lesion regions are concentrated in the left prefrontal cortex, striatum, and cerebellum. Among them, the striatum and cerebellum exhibit higher occurrence probabilities, whereas prefrontal lesion regions appear less frequently but remain clearly detectable. This pattern is consistent with the sub-dataset analysis and further supports the prefrontal–striatal–cerebellar circuit model of ADHD [43]. Abnormal communication within this circuit has been widely regarded as a core neural mechanism underlying deficits in attention, reward processing, and executive control in ADHD [44].

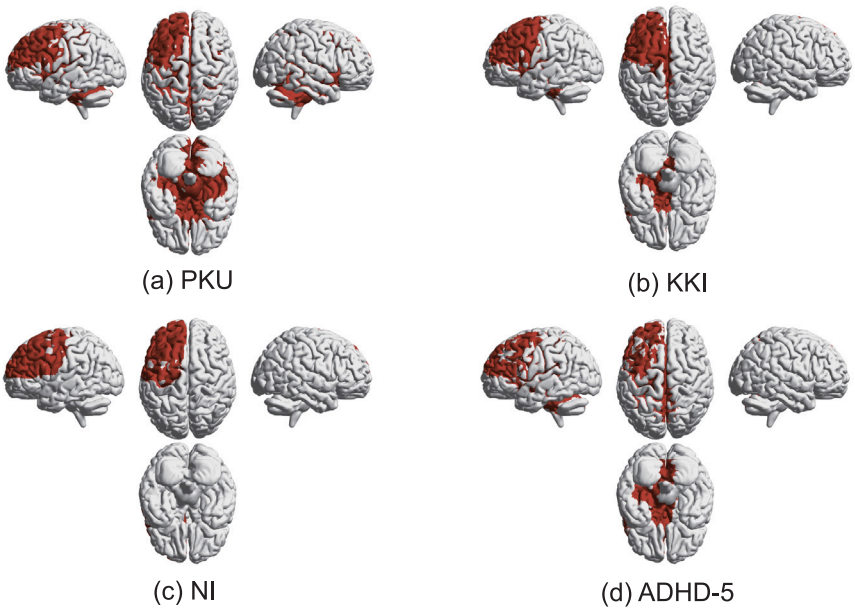

Fig. 4. Visualization of lesion regions detected by the proposed model in different sub-datasets: (a) PKU, (b) KKI, (c) NI, and (d) ADHD-5.

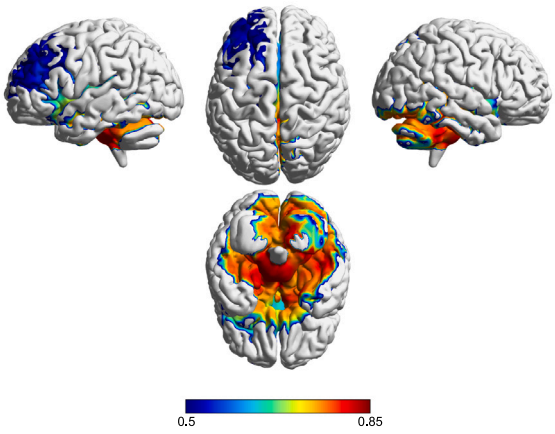

Fig. 5. Probability distribution of lesion regions in the ADHD-200 dataset.

6.5.3. Cross-group discussion

The subgroup visualizations indicate that the lesion regions identified by the proposed framework are modulated by both age and sex. As shown in Fig. 6, the Child subgroup exhibits more concentrated and reproducible lesion regions in frontal, ventral, and deep-brain areas, whereas the Adolescent subgroup shows a more diffuse pattern with persistent ventral frontal and limbic-related involvement. This suggests that ADHD-related abnormalities in younger subjects are more strongly anchored in fronto-striatal and attention-related systems, while in the older subgroup the lesion regions become relatively more distributed and more closely associated with default mode network (DMN)-limbic interactions.

This age-related dissociation is consistent with prior neuroimaging findings. Guo et al. reported that childhood ADHD is more strongly characterized by abnormal coupling between the somatomotor and dorsal attention networks, whereas older subjects show more distinct DMN-limbic abnormalities [45]. Hoogman et al. further showed that subcortical alterations in ADHD are more pronounced in children and attenuate with age [46].

A comparable dissociation is observed in the sex-based comparison. As shown in Fig. 7, the Female subgroup presents broader lesion regions in ventral and medial frontal areas, especially around orbitofrontal and ventromedial prefrontal territories, whereas the Male subgroup shows a more focal pattern in dorsal and lateral frontal regions, with additional involvement of inferior frontal, premotor, and striatal-related areas. These findings suggest that female ADHD is more ventral-frontal and limbic-oriented, whereas male ADHD is more strongly linked to frontal inhibitory-control and motor-related circuitry.

This interpretation is also supported by previous studies. Jacobson et al. reported a sex-by-diagnosis dissociation in white matter microstructure, showing reduced fractional anisotropy in bilateral primary motor regions in boys with ADHD, but higher fractional anisotropy in bilateral medial orbitofrontal cortex in girls with ADHD [47]. Rosch et al. further showed that girls with ADHD exhibit more pronounced fronto-subcortical intrinsic connectivity abnormalities, whereas boys present a pattern more compatible with classical inhibitory-control dysfunction [48].

Overall, the cross-group results suggest that the proposed PH-based framework captures not only shared ADHD-related abnormalities, but also biologically meaningful subgroup specific variations. In particular, childhood ADHD appears more strongly linked to fronto-striatal, sensorimotor, and attention-control abnormalities, whereas the older subgroup shows relatively greater DMN-limbic involvement. Similarly, female ADHD is more ventral-frontal and limbic-oriented, whereas male ADHD is more dominated by frontal inhibitory-control and motor-related systems.

It is worth noting that although the identified lesion regions are consistent with previously reported ADHD-related circuits, the current localization analysis remains mainly qualitative. Since voxel-level ground-truth lesion annotations are generally unavailable for ADHD, direct quantitative validation is challenging. Therefore, the detected lesion regions should be interpreted as model-derived discriminative regions rather than confirmed pathological lesions. Future work will incorporate atlas-based regional mapping, bootstrap stability analysis, and statistical testing to further validate the reliability of localization results.

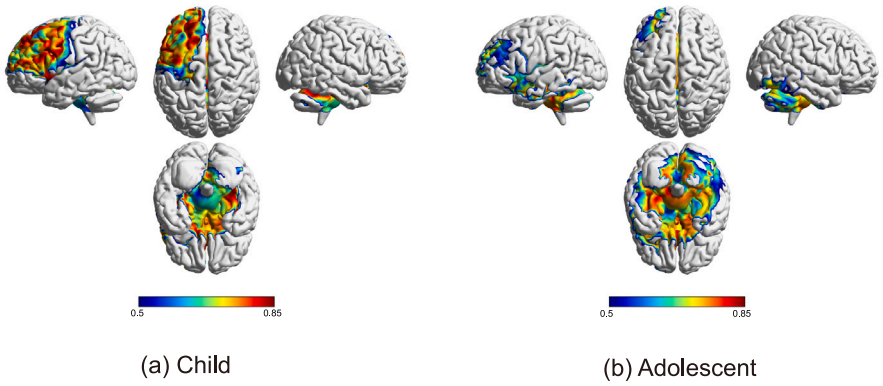

Fig. 6. BrainNet visualizations of subgroup-level ADHD probability decision matrices for the age-based analysis. Warmer colors indicate higher occurrence probabilities of lesion regions among ADHD subjects. (a) Child subgroup. (b) Adolescent subgroup. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

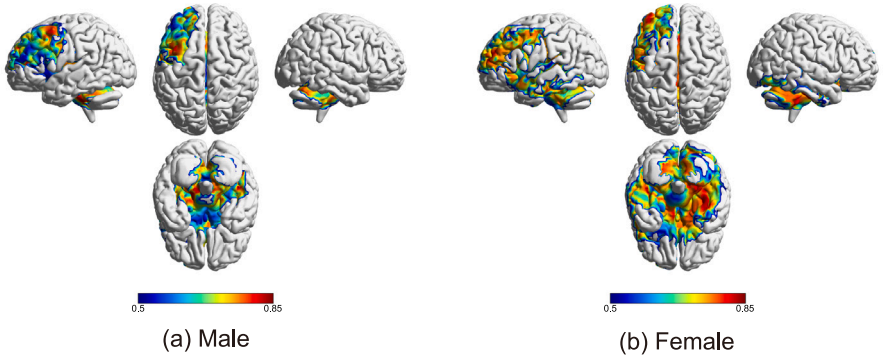

Fig. 7. BrainNet visualizations of subgroup-level ADHD probability decision matrices for the sex-based analysis. Warmer colors indicate higher occurrence probabilities of lesion regions among ADHD subjects. (a) Male subgroup. (b) Female subgroup. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Conclusion

This study proposes an explainable MRI-based framework for ADHD classification that integrates deep learning with PH to enable diagnosis with explicit spatial interpretability. Evaluated on the ADHD-200 dataset, the framework achieves competitive performance and identifies discriminative lesion regions mainly distributed in the prefrontal–striatal–cerebellar circuitry. It also reveals potentially meaningful spatial heterogeneity across age and sex subgroups. Although the current implementation uses a ResNet-based slice classifier, the framework is readily compatible with foundation-model-based medical image encoders. In future work, self-supervised MRI foundation models could provide stronger slice-level representations, while the PH-based stage could serve as an interpretable reasoning and spatial recovery module. In this sense, the proposed method offers a topology-driven framework that can be naturally extended to future medical foundation models.

However, existing studies still face several challenges. First, current frameworks primarily rely on traditional ResNet backbones and single-modal slice-level processing, which may limit the comprehensive exploration of complex 3D anatomical structures and brain connectivity. To address this, future research will focus on upgrading the feature extraction module by introducing advanced topological graph neural networks. Furthermore, we aim to overcome the current limitation of manual hyperparameter tuning for topological features by incorporating Differentiable PH. By parameterizing topological descriptors and embedding them directly into the neural network's loss function, the model will enable the automated optimization of topological parameters via backpropagation. This shift towards an end-to-end joint training framework will effectively bridge slice-level feature learning with subject-level topological classification, thereby enhancing the model's generalization and adaptability across diverse clinical cohorts.

CRediT authorship contribution statement

Peng Wang: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Jiayi Duan:** Investigation, Formal analysis, Data curation. **Yuqing Xing:** Writing – review & editing, Methodology. **Ruihang Xu:** Writing – review & editing, Methodology, Conceptualization. **Anyuan Xu:** Methodology, Data curation. **Haodong Chen:** Data curation. **Shengchao Hu:** Writing – review & editing, Visualization, Methodology, Data curation. **Tao Wang:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Shuang Liu:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62376187, the Science Fund for Distinguished Young Scholars of Tianjin under Grant 23JCJQJC00060, and the Autonomous Project of Haihe Laboratory of Brain-Computer Interaction and Human–Machine Integration under Grant 25HHNJSS00004.

Data availability

Data will be made available on request.

References

[1] X. Han, M. Lei, J. Li, Hypergraph-based semantic and topological self-supervised learning for brain disease diagnosis, Pattern Recognit. 169 (2026) 111921.

[2] K. Ma, Q. Zhu, X. Wen, X. Yang, D. Zhang, Sliced Wasserstein graph kernel for measuring global topological similarity of brain functional networks, Pattern Recognit. (2025) 112208.

[3] Y. Mizuno, M. Yamashita, Q. Shou, S. Hamatani, W. Cai, A brief review of MRI studies in patients with attention-deficit/hyperactivity disorder and future perspectives, Brain Dev. 47 (2) (2025) 104340.

[4] M.T. Vlaardingerbroek, J.A. Boer, Magnetic Resonance Imaging: Theory and Practice, Springer Science & Business Media, 2013.

[5] X. Wu, W. Wu, X. Zhang, J. Zhang, Adaptive and asynchronous integration of gray and white matter fMRI for brain disorder diagnosis, Pattern Recognit. (2026) 113502.

[6] V. Kulkarni, B. Nemade, S. Patel, K. Patel, S. Velpula, A short report on ADHD detection using convolutional neural networks, Front. Psychiatry 15 (2024) 1426155.

[7] T. Wang, H. Lu, J. Duan, T. Meng, R. Mao, S. Liu, D. Ming, Explainable affective body expression recognition with multi-scale spatiotemporal encoding and LLM-based reasoning, IEEE Trans. Affect. Comput. (2026).

[8] T. Wang, R. Mao, S. Liu, E. Cambria, D. Ming, Explainable multi-frequency and multi-region fusion model for affective brain-computer interfaces, Inf. Fusion 118 (2025) 102971.

[9] T. Wang, S. Liu, F. He, W. Dai, M. Du, Y. Ke, D. Ming, Emotion recognition from full-body motion using multiscale spatio-temporal network, IEEE Trans. Affect. Comput. 15 (3) (2023) 898–912.

[10] T. Wang, S. Liu, F. He, M. Du, W. Dai, Y. Ke, D. Ming, Affective body expression recognition framework based on temporal and spatial fusion features, Knowl.-Based Syst. 308 (2025) 112744.

[11] G. Carlsson, Topology and data, Bull. Am. Math. Soc. 46 (2) (2009) 255–308.

[12] A. Zomorodian, G. Carlsson, Computing persistent homology, in: Proceedings of the Twentieth Annual Symposium on Computational Geometry, 2004, pp. 347–356.

[13] Z. Cang, G.-W. Wei, TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, PLoS Comput. Biol. 13 (7) (2017) e1005690.

[14] G. Polanczyk, et al., The worldwide prevalence of ADHD: a systematic review and metaregression analysis, Am. J. Psychiatry 164 (6) (2007) 942–948.

[15] F. Li, Y. Cui, Y. Li, L. Guo, X. Ke, J. Liu, X. Luo, Y. Zheng, J.F. Leckman, Prevalence of mental disorders in school children and adolescents in China: diagnostic data from detailed clinical assessments of 17,524 individuals, J. Child Psychol. Psychiatry 63 (1) (2022) 34–46.

[16] S. Liu, L. Zhao, J. Zhao, B. Li, S.-H. Wang, Attention deficit/hyperactivity disorder classification based on deep spatio-temporal features of functional magnetic resonance imaging, Biomed. Signal Process. Control 71 (2022) 103239.

[17] T. Zhang, C. Li, P. Li, Y. Peng, X. Kang, C. Jiang, F. Li, X. Zhu, D. Yao, B. Biswal, et al., Separated channel attention convolutional neural network (SC-CNN-attention) to identify ADHD in multi-site rs-fMRI dataset, Entropy 22 (8) (2020) 893.

[18] A.M. Aradhya, V. Subbaraju, S. Sundaram, N. Sundararajan, Discriminant Spatial Filtering Method (DSFM) for the identification and analysis of abnormal resting state brain activities, Expert Syst. Appl. 181 (2021) 115074.

[19] D.C. Lohani, B. Rana, ADHD diagnosis using structural brain MRI and personal characteristic data with machine learning framework, Psychiatry Res.: Neuroimaging 334 (2023) 111689.

[20] Edelsbrunner, Letscher, Zomorodian, Topological persistence and simplification, Discrete Comput. Geom. 28 (2002) 511–533.

[21] G. Carlsson, A. Zomorodian, A. Collins, L. Guibas, Persistence barcodes for shapes, in: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, 2004, pp. 124–135.

[22] A. Bukkuri, N. Andor, I.K. Darcy, Applications of topological data analysis in oncology, Front. Artif. Intell. 4 (2021) 659037.

[23] E.N. Pitsik, V.A. Maximenko, S.A. Kurkin, A.P. Sergeev, D. Stoyanov, R. Paunova, S. Kandilarova, D. Simeonova, A.E. Hramov, The topology of fMRI-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder, Chaos Solitons Fractals 167 (2023) 113041.

[24] W. Zhang, S. Xia, X. Tang, X. Zhang, D. Liang, Y. Wang, Topological analysis of functional connectivity in Parkinson's disease, Front. Neurosci. 17 (2023) 1236128.

[25] A. François, R. Tinarrage, Train-free segmentation in MRI with cubical persistent homology, 2024, arXiv preprint arXiv:2401.01160.

[26] H. Edelsbrunner, J.L. Harer, Computational Topology: An Introduction, American Mathematical Society, 2022.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[28] R. Ghrist, Barcodes: the persistent topology of data, Bull. Am. Math. Soc. 45 (1) (2008) 61–75.

[29] A. consortium, The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience, Front. Syst. Neurosci. 6 (2012) 62.

[30] M. Harm, M. Hope, A. Household, American psychiatric association, 2013, diagnostic and statistical manual of mental disorders, 5th edn, washington, dc: American psychiatric association anderson, j, sapey, b, spandler, h (eds.), 2012, distress or disability?, lancaster: Centre for disability research, Arya 347 (2013) 64.

[31] M. Xia, J. Wang, Y. He, BrainNet viewer: a network visualization tool for human brain connectomics, PLoS One 8 (7) (2013) e68910.

[32] S. Farzi, S. Kianian, I. Rastkhadive, Diagnosis of attention deficit hyperactivity disorder using deep belief network based on greedy approach, in: 2017 5th International Symposium on Computational and Business Intelligence, ISCBI, IEEE, 2017, pp. 96–99.

[33] L. Zou, J. Zheng, C. Miao, M.J. Mckeown, Z.J. Wang, 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI, IEEE Access 5 (2017) 23626–23636.

[34] A. Riaz, et al., Deep fMRI: An end-to-end deep network for classification of fMRI data, in: 2018 Ieee 15th International Symposium on Biomedical Imaging, Isbi 2018, IEEE, 2018, pp. 1419–1422.

[35] N.A. Khan, S.A. Waheeb, A. Riaz, X. Shang, A novel knowledge distillation-based feature selection for the classification of ADHD, Biomolecules 11 (8) (2021) 1093.

[36] P. Wang, X. Zhao, J. Zhong, Y. Zhou, Localization and diagnosis of attention-deficit/hyperactivity disorder, Healthcare 9 (4) (2021) 372.

[37] J. Li, G. Liu, J. Ji, Pre-training and fine-tuning transformer for brain network classification, in: 2024 IEEE International Conference on Medical Artificial Intelligence, MedAI, IEEE, 2024, pp. 136–144.

[38] G. Bush, P. Luu, M.I. Posner, Cognitive and emotional influences in anterior cingulate cortex, Trends Cogn. Sci. 4 (6) (2000) 215–222.

[39] A. Sebastian, P. Jung, A. Krause-Utz, K. Lieb, C. Schmahl, O. Tüscher, Frontal dysfunctions of impulse control–a systematic review in borderline personality disorder and attention-deficit/hyperactivity disorder, Front. Hum. Neurosci. 8 (2014) 698.

[40] G.C. Burgess, B.E. Depue, L. Ruzic, E.G. Willcutt, Y.P. Du, M.T. Banich, Attentional control activation relates to working memory in attention-deficit/hyperactivity disorder, Biol. Psychiatry 67 (7) (2010) 632–640.

[41] E.M. Valera, S.V. Faraone, K.E. Murray, L.J. Seidman, Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder, Biol. Psychiatry 61 (12) (2007) 1361–1369.

[42] F.X. Castellanos, P.P. Lee, W. Sharp, N.O. Jeffries, D.K. Greenstein, L.S. Clasen, J.D. Blumenthal, R.S. James, C.L. Ebens, J.M. Walter, et al., Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder, Jama 288 (14) (2002) 1740–1748.

[43] F.X. Castellanos, E.J. Sonuga-Barke, M.P. Milham, R. Tannock, Characterizing cognition in ADHD: beyond executive dysfunction, Trends Cogn. Sci. 10 (3) (2006) 117–123.

[44] Y. Feng, D. Zhi, Y. Zhu, X. Guo, X. Luo, C. Dang, L. Liu, J. Sui, L. Sun, Symptom-guided multimodal neuroimage fusion patterns in children with attention-deficit/hyperactivity disorder and its potential "brain structure–function-cognition–behavior" pathological pathways, Eur. Child Adolesc. Psychiatry 33 (7) (2024) 2141–2152.

[45] X. Guo, D. Yao, Q. Cao, L. Liu, Q. Zhao, H. Li, F. Huang, Y. Wang, Q. Qian, Y. Wang, et al., Shared and distinct resting functional connectivity in children and adults with attention-deficit/hyperactivity disorder, Transl. Psychiatry 10 (1) (2020) 65.

[46] M. Hoogman, J. Bralten, D.P. Hibar, M. Mennes, M.P. Zwiers, L.S. Schweren, K.J. van Hulzen, S.E. Medland, E. Shumskaya, N. Jahanshad, et al., Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis, Lancet Psychiatry 4 (4) (2017) 310–319.

[47] L.A. Jacobson, D.J. Peterson, K.S. Rosch, D. Crocetti, S. Mori, S.H. Mostofsky, Sex-based dissociation of white matter microstructure in children with attention-deficit/hyperactivity disorder, J. Am. Acad. Child Adolesc. Psychiatry 54 (11) (2015) 938–946.

[48] K.S. Rosch, S.H. Mostofsky, M.B. Nebel, ADHD-related sex differences in fronto-subcortical intrinsic functional connectivity and associations with delay discounting, J. Neurodev. Disord. 10 (1) (2018) 34.