Neuroimage: Reports 6 (2026) 100338

Contents lists available at ScienceDirect

Neuroimage: Reports

journal homepage: www.sciencedirect.com/journal/neuroimage-reports

An interpretable cross-attentive multi-modal MRI fusion framework for schizophrenia identification

Ziyu Zhou [a], Anton Orlichenko [b], Gang Qu [b], Zening Fu [e], Zhengming Ding [a,*], Julia Stephen [c], Tony Wilson [d], Vince Calhoun [e], Yu-Ping Wang [b,*]

[a] Tulane University, Department of Computer Science, 6823 St. Charles Ave, New Orleans, 70118, LA, USA
[b] Tulane University, Department of Biomedical Engineering, 6823 St. Charles Ave, New Orleans, 70118, LA, USA
[c] Mind Research Network, 1101 Yale Blvd NE, Albuquerque, 87106, NM, USA
[d] Boys Town National Research Hospital, Institute for Human Neuroscience, 14000 Boys Town Hospital Rd, Boys Town, 68010, NE, USA
[e] Georgia State University, Georgia Institute of Technology, Emory University, Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), 55 Park Pl NE, Atlanta, 30303, GA, USA

ARTICLE INFO

Dataset link: https://nda.nih.gov/edit_collection.html?id=2274, https://coins.trendscenter.org/, https://www.nitrc.org/projects/fbirn/

Keywords:
Multi-modal MRI
Cross-modal attention
Transformer
Schizophrenia

ABSTRACT

Functional MRI (fMRI) and structural MRI (sMRI) offer complementary insights into brain function and anatomy, but their integration for schizophrenia identification remains challenging due to modality heterogeneity. Many existing methods fall short of effective modeling of the interaction between two modalities. We propose CAMF, a Cross-Attentive Multi-modal Fusion framework that employs self-attention to capture intra-modal patterns and cross-attention to learn inter-modal relationships. In addition, we introduce the gradient-guided score-class activation map to enhance interpretability by highlighting salient features. Our approach significantly improves the accuracy in classifying schizophrenia, as demonstrated by the evaluation of multi-modal brain imaging datasets from four cohorts of schizophrenia studies. Furthermore, the model identifies functional networks and anatomical regions aligned with established biomarkers. CAMF provides an accurate and interpretable framework for multimodal brain imaging analysis, offering new insights into schizophrenia-related alterations.

1. Introduction

Medical imaging powered by machine learning has been widely employed for the identification of mental disorders such as schizophrenia (Oh et al., 2020, Meng et al., 2023), autism (Katuwal et al., 2015), and Alzheimer's disease (AD) (Tomassini et al., 2021, Gao et al., 2023). Among imaging modalities, functional magnetic resonance imaging (fMRI) and structural magnetic resonance imaging (sMRI) have emerged as popular imaging modalities for mental disorder analysis. fMRI offers insights into the functional organization of the brain by quantifying changes in the blood-oxygen-level-dependent (BOLD) signal of the human brain. The functional connectivity (FC) derived from fMRI depicts the correlation between brain regions of interest (ROIs) and has proven effective in tasks such as age prediction (Li et al., 2018) and disease identification (Zhang et al., 2017), and even serves as a "brain fingerprint" reflecting individual cognitive and behavioral traits (Wang et al., 2021; Finn et al., 2015, Yan et al., 2024). sMRI, on the other hand, provides anatomical features such as cortical thickness and surface area, which have been linked to neurodevelopmental and disease phenotypes.

The analysis on fMRI and sMRI is promising to bridge the gab between clinical findings and the development of theories for the underlying mechanism of human brain, as demonstrated by prior research (Chen et al., 2024). For example, a study (Liu et al., 2025) analyzed the fMRI and sMRI scans of adolescent subjects and evaluated the associations between proportion of prosocial/delinquent friends and the structural and functional architecture of the brain, cognition, as well as behavioral and emotional dysregulation. Similarly, the same adolescent cohort was used to discover subtypes of anxiety-impulsivity in preadolescent internalizing disorders (Fan et al., 2023). Another study (Wu et al., 2025) mapped the developmental trajectory of structural brain asymmetry and discovered how the developing brain asymmetry shapes cognitive and psychiatric outcomes in adolescence.

* Corresponding authors.
E-mail addresses: zzhou11@tulane.edu (Z. Zhou), aorlichenko@tulane.edu (A. Orlichenko), gqu1@tulane.edu (G. Qu), zfu@gsu.edu (Z. Fu), zding1@tulane.edu (Z. Ding), jstephen@mrn.org (J. Stephen), tony.wilson@boystown.org (T. Wilson), vcalhoun@gsu.edu (V. Calhoun), wyp@tulane.edu (Y.-P. Wang).

https://doi.org/10.1016/j.ynirp.2026.100338
Received 12 October 2025; Received in revised form 7 March 2026; Accepted 16 March 2026

While integrating or fusing fMRI and sMRI may enable a more comprehensive view of brain pathology by leveraging both functional and structural markers (Calhoun and Sui, 2016), the efficient integration of sMRI and fMRI remains challenging, largely due to the heterogeneity of these two modalities. Various methods have been proposed to integrate heterogeneous data from multiple modalities. Multi-set canonical correlation analysis (MCCA) (Sui et al., 2013) was introduced for the fusion of multi-modality brain image data including fMRI and sMRI. Mousavian et al. (2021) concatenated the structural and functional similarity matrices extracted from sMRI and fMRI to enhance the multi-modal classification of major depressive disorder. Hojjati et al. (2019) also integrated the fMRI and gray matter and white matter features extracted from sMRI for Alzheimer's disease identification via concatenation (Hojjati et al., 2019). However, these integration methods fail to fully utilize the potential synergy between these modalities by using simple concatenation, being unable to explore the interaction between multiple modalities.

Multi-modal deep learning models have therefore emerged as effective solutions for integrating diverse brain imaging modalities. For example, Zu et al. (2016) introduced a label-aligned regularization term in a multi-kernel learning framework to utilize the relationships between multiple modalities for the selection of important feature subsets resulting in improved AD classification. Another method (Yang et al., 2023) utilized graph-based networks to fuse the structural and functional connectivity extracted from DTI and fMRI, achieving promising performance. Other work (Khalilullah et al., 2023) proposed the parallel multilink joint ICA method to integrate intrinsic connectivity networks (ICNs) extracted from fMRI and gray matter features. Liu's team (Liu et al., 2022) integrated brain functional and structural graphs constructed from fMRI and sMRI respectively and improved the identification of neuropsychiatric disorders. Although these methods can explore the interaction between modalities, they rely on high-level features such as functional and structural connectivity extracted from MRI, where the interpretability of nonlinear data integration still needs investigation.

The transformer model (Vaswani et al., 2017) has explored the attention mechanism to extract interactions for sequence data. This mechanism shows great potential of data integration for our task in terms of fusing latent features from fMRI and sMRI, as it can discover interrelations in heterogeneous data across modalities (Qu et al., 2023). To enhance interpretability and capture structural-functional complementarity, we propose a two-level data fusion framework, Cross-Attentive Multi-modal Fusion (CAMF), incorporating the attention mechanism for schizophrenia identification. At the first level, we use self-attention (SA) modules to extract interactions within each modality and cross-attention (CA) modules to explore interactions between fMRI and sMRI (Zhu et al., 2022). At the second level, we integrate embeddings from these four attention pathways, dynamically updating pathway weights throughout the model training process to achieve optimal integration.

To facilitate model interpretability, we design a gradient-based Score-CAM to maximize the capability of both gradient-based and perturbation-based interpretation methods. The gradient-based interpretation is only applicable to end-to-end models and requires the calculation of gradients. It assesses input influence through network backpropagation. Conversely, perturbation-based methods are general to any model. Such methods alter inputs to observe output variations, which interprets the result at the feature level. However, the sensitivity depends on the evaluation metrics. Score-CAM (Wang et al., 2020) interprets the model by generating the class activation mapping with a combination of perturbation-based scores and feature maps, targeting both sensitivity and generalizability. In our study, we further advance the Score-CAM with gradient guidance to generate more precise saliency maps based on both fMRI and sMRI. To be specific, we combine gradient and score-CAM saliency maps to benefit from both gradient-based and perturbation-based methods. The disease-related brain regions identified by both modalities corroborate findings from previous studies, further validating the reliability and interpretability of our model.

The rest of the paper is organized as follows: Section 2 gives an overview of the proposed framework. Section 3 contains experiments on various datasets, comparing its performance and interpretability with other multi-modal fusion methods. Section 4 examines schizophrenia-related brain functional networks and structural regions identified by our framework, while also discussing its limitations and outlining potential avenues for future scope. Section 5 summarizes our main contributions and findings.

2. Methodology

2.1. Preliminary

Each subject has both fMRI and sMRI data. For fMRI modality, we use the functional connectivity (FC) matrix as input, denoted as $\mathbf{X}_i^f \in \mathbb{R}^{264\times264}$ where i is the index of the sample, and 264 is the number of region of interest (ROI) based on power parcellation scheme (Power et al., 2011). For the sMRI modality, we directly used the 3D voxel-level data as the input, denoted as $\mathbf{X}_i^s \in \mathbb{R}^{121\times145\times121}$, with the specific dimensions in our data described in Section 3.1.1. Thus, the data used in our classification are $\mathcal{D}_{tr} = \{\mathbf{X}_i^f, \mathbf{X}_i^s, Y_i\}_{i=1}^N$ where N is the sample size. For this binary classification task, the output Y_i is a one-hot vector, with an entry for healthy controls (HC) and schizophrenia subjects (SZ), respectively.

2.2. Cross-attentive multi-modal fusion

2.2.1. Framework overview

The overall architecture of our proposed framework is shown in Fig. 1. Specifically, a subject-specific 2D FC derived from fMRI is used as the input to a 2D convolution neural network (CNN), where each entry of FC represents the connection between a pair of ROIs. For sMRI, voxel-level structural MRI data are used as the input to a 3D CNN.

Our proposed framework employed CNNs as feature extractors since they are ideally suited for handling high-dimensional 2D/3D data such as fMRI FC matrices and the volumetric data from sMRI. The FC derived from fMRI is a symmetric 2D matrix, where the order of rows and columns reflects the spatial relationship between ROIs. A global pooling layer is applied to compress the feature map of each channel and generate vectorized latent features, which are denoted as:

$$\mathbf{f}_1 = \mathbf{GP}(\mathbf{CNN_{2D}}(\mathbf{X}^f)), \tag{1}$$

where $\mathbf{f}_1 \in \mathbb{R}^{d_1}$ with the hyperparameter $d_1 = 256$ being the number of channels of the last convolutional layer; and $\mathbf{GP}(\cdot)$ represents the global pooling layer. For the sMRI modality, we utilize voxel-level images that contain anatomically-related information. A 3D CNN is used as the backbone and complemented by a global pooling layer, producing the following features:

$$\mathbf{f}_2 = \mathbf{GP}(\mathbf{CNN_{3D}}(\mathbf{X}^s)), \tag{2}$$

where $\mathbf{f}_2 \in \mathbb{R}^{d_1}$. In this paper, we use global max pooling after the last convolutional layer to compress the feature maps of each channel.

2.2.2. Intra- and inter-modality interaction

To capture both intra-modal and inter-modal interactions, we utilize two fusion mechanisms, each integrating four pathways of attention modules. This strategy allows us to incorporate the interactions between two modalities while uncovering more intricate interactions among ROIs with improved interpretability.

The two self-attention (SA) modules focus on exploring interactions among subjects within individual modalities, i.e., revealing intra-modal relationships among subjects. Conversely, the cross-attention (CA) modules are tailored to explore the interactions between the two different modalities, thereby addressing the inter-modal interactions.

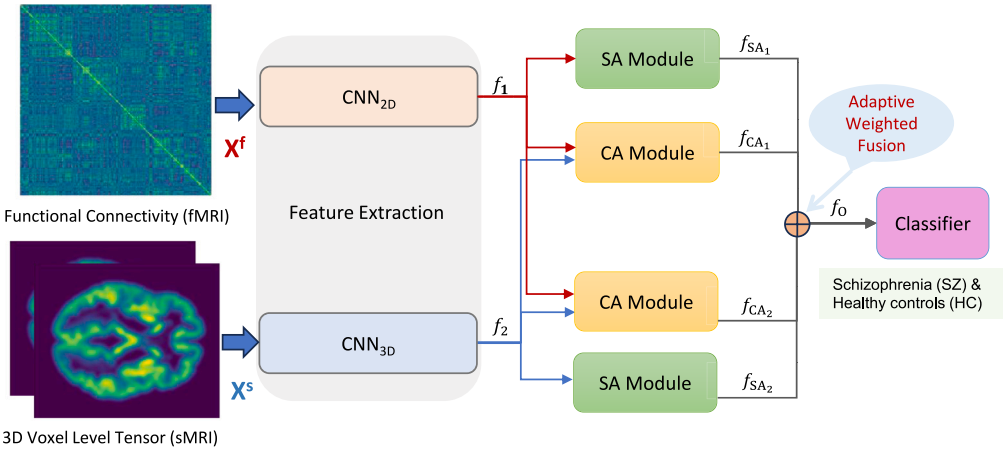

Fig. 1. An overview of the proposed framework. The backbones consist of two CNN modules to extract features from the fMRI and sMRI data. Then two self-attention (SA) modules and two cross-attention (CA) modules fuse the features at the first level. The latent features are then combined by the optimal weights and input to a classifier.

In each SA module, Query $\boldsymbol{q}_i = \boldsymbol{f}_i W_{q_i}$, Key $\boldsymbol{k}_i = \boldsymbol{f}_i W_{k_i}$ and Value $\boldsymbol{v}_i = \boldsymbol{f}_i W_{v_i}$ are generated from the same latent feature $\boldsymbol{f}_i, i \in \{1,2\}$, where $W_{q_i}, W_{k_i}, W_{v_i} \in \mathbb{R}^{d_1 \times d_2}$ are learnable parameters and $\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i \in \mathbb{R}^{n \times d_2}$. In our experiments, we set the hyperparameter $d_2 = 256$. Then the output modality-specific feature is

$$\boldsymbol{f}_{\mathrm{SA}_i} = \mathrm{softmax}(\frac{\boldsymbol{q}_i \boldsymbol{k}_i^\top}{\sqrt{d_2}})\boldsymbol{v}_i. \tag{3}$$

For the cross-modal fusion scenario, we learn the interaction features between both modalities in the following way:

$$\boldsymbol{f}_{\mathrm{CA}_1} = \mathrm{softmax}(\frac{\mathbf{q}_1 \boldsymbol{k}_2^\top}{\sqrt{d_2}})\boldsymbol{v}_2, \tag{4}$$

$$\boldsymbol{f}_{\mathrm{CA}_2} = \mathrm{softmax}(\frac{\mathbf{q}_2 \boldsymbol{k}_1^\top}{\sqrt{d_2}})\boldsymbol{v}_1. \tag{5}$$

In the second CA module, we reverse the order of modalities to address the issue of asymmetry, enabling a more effective exploration of the interaction between the two modalities.

To fuse the four outputs generated from the previous interaction module, we investigate an adaptive fusion strategy that automatically identifies the contribution of each pathway. The output features from the four attention modules $\boldsymbol{f}_{\mathrm{SA}_1}, \boldsymbol{f}_{\mathrm{SA}_2}, \boldsymbol{f}_{\mathrm{CA}_1}, \boldsymbol{f}_{\mathrm{CA}_2} \in \mathbb{R}^d$ are subsequently fused by weighted sum

$$\boldsymbol{f}_O = \alpha_1 \boldsymbol{f}_{\mathrm{SA}_1} + \alpha_2 \boldsymbol{f}_{\mathrm{SA}_2} + \alpha_3 \boldsymbol{f}_{\mathrm{CA}_1} + \alpha_4 \boldsymbol{f}_{\mathrm{CA}_2}, \tag{6}$$

where α_{1-4} are learnable parameters that sum to be 1, achieved through the $\mathrm{softmax}(\cdot)$ operation. This fusion process enables the model to learn an optimal combination of the four attention pathways. The fused feature $\boldsymbol{f}_O$ is then input to the classifier.

Given the fused feature $\boldsymbol{f}_O$, a final prediction vector with two entries is calculated using a multilayer perceptron (MLP) classifier $C(\cdot)$. Then we deploy the cross-entropy loss Eq. (7) to guide model training with ground-truth supervision $\mathbf{Y}$.

$$\ell_{ce} = \mathrm{CrossEntropy}(C(\boldsymbol{f}_O), \mathbf{Y}). \tag{7}$$

We employed the Adam optimizer with a weight decay rate of 1e-4 to optimize the model parameters. We adopted the initialization proposed by Kaiming described in He et al. (2015) to initialize the model parameters.

Table 1
Statistics of the datasets.

Dataset	SZ	HC	Males	Females
COBRE	57	81	109	29
FBIRN	127	152	208	71
MPRC	80	123	122	81
BSNIP	199	243	239	203
Total	463	599	678	314

3. Experiments

3.1. Dataset

3.1.1. Combined dataset

The evaluation was first assessed using the samples combined from three datasets: Centers of Biomedical Research Excellence (COBRE) (Aine et al., 2017), The Function Biomedical Informatics Research Network (FBIRN) (Keator et al., 2016) and Maryland Psychiatric Research Center (MPRC) (Adhikari et al., 2019). Together, these datasets comprise 356 HC samples and 264 SZ samples. Table 1 lists the detailed sample size of each subset.

3.1.2. Bipolar and Schizophrenia network for intermediate phenotypes (BSNIP)

We conducted independent testing on datasets with different sources and sampling protocols for the generalizability of the framework. Herein, we used the Bipolar and Schizophrenia Network for Intermediate Phenotypes (BSNIP) dataset (Tamminga et al., 2014) to validate the performance. Our BSNIP dataset includes 243 HC samples and 199 SZ samples.

3.1.3. Data preprocessing

We followed the same standard pipeline for MRI data pre-processing as described in Rahaman et al. (2023). Specifically, the sMRI scans were preprocessed using the statistical parametric mapping (SPM12) toolbox.[1] The unified segmentation and normalization were applied to the sMRI scans for gray matter, white matter, and cerebrospinal fluid (CSF), and a spatial normalization algorithm was used to generate modulated gray matter volume (GMV) maps and unmodulated gray matter density (GMD) maps. Then the GMV and GMD were smoothed

[1] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

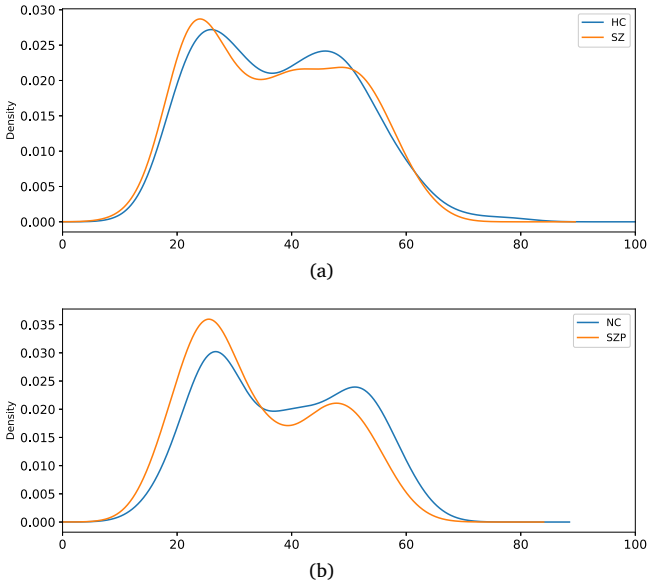

Fig. 2. Age distribution across datasets: (a) For the combined dataset from COBRE, FBIRN, and MPRC; (b) For the BSNIP dataset.

using a Gaussian kernel with a full width at half maximum (FWHM) = 6 mm.

The fMRI data were preprocessed using the SPM toolbox within MATLAB 2020b. The first five scans were removed for the signal equilibrium and participants' adaptation to the scanner's noise. We performed rigid body motion correction using the toolbox in SPM to correct subject head motion, followed by the slice-timing correction to account for timing differences in slice acquisition. The fMRI data were subsequently normalized into the standard Montreal Neurological Institute (MNI) space using an echo-planar imaging (EPI) template and were resampled to $3 \times 3 \times 3\ mm^3$ isotropic voxels. The resampled fMRI images were further smoothed using a Gaussian kernel with a FWHM = 6 mm. We then applied the Power atlas (Power et al., 2011) at the voxel-level image, producing time-series data across 264 regions of interest (ROIs). Then we derived the functional connectivity (FC) matrix via Pearson correlation between the 264 ROIs.

We further investigated the potential impact of confounders such as gender, age, and head motion. Despite gender disparities in our datasets, evidence (Miller et al., 2011) exists to consistently show that the risk of developing schizophrenia is not significantly influenced by an individual's gender, underscoring that gender plays a negligible role in this risk factor. Fig. 2 is the age distribution of HCs and SZs in each dataset, indicating an unbiased age distribution relative to disease classification. In alignment with our previous research (Qu et al., 2021b), we tested the hypothesis that there is no relationship ($p < 0.05$) between the disease groups and the mean framewise displacement (FD) (Power et al., 2012) using the matrix of Pearson correlation coefficients.

3.2. Comparison experiment

3.2.1. Cross validation

In this study, we adopted 5-fold cross-validation to evaluate the model performance. The patient index codes were shuffled and divided evenly into 5 folds, and one fold was left out as the test set. For the remaining 4 folds, we selected 1/8 to be the validation set and the larger portion to be the training set. The averaged results of 5 folds are presented for all experiments.

3.2.2. Evaluation metrics

An potential concern with the combined dataset was the imbalance in disease class. To address this concern, we employed additional evaluation metrics, such as F1 score and Matthew's correlation coefficient (MCC), alongside the standard measure of classification accuracy.

The F1 score is calculated as:

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN}, \tag{8}$$

where the minimum value is 0 and the maximum value is +1.

The MCC is defined as

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}, \tag{9}$$

where the minimum value is −1 and the maximum value is +1. MCC is recommended by the National Institutes of Health (NIH) (Chicco and Jurman, 2020) to be a reliable statistic because it produces a high score when the prediction obtained good results in all of the four confusion matrix categories (true positives: TP, false negatives: FN, true negatives: TN, and false positives: FP). In addition, we also report the class-weighted cross entropy loss, AUCROC, and the average precision score (Ave PRC) for the ablation studies.

3.2.3. Baselines

We compared the combination of different feature extractors, fusion methods, and classifiers, which demonstrated the superiority of our proposed model.

For feature extractors, we compared principal component analysis (PCA) and CNNs. For PCA, we vectorize the inputs for both modalities, and apply PCA to reduce the dimension to $d = 256$, matching our framework's latent feature size.

Regarding fusion methods, we explored element-wise sum operation, concatenation, and our proposed CAMF; for classifiers, we adopted conventional models such as Support Vector Machine (SVM) (Cortes and Vapnik, 1995), linear SVM with stochastic gradient descent learning (SGDClassifier) (Amari, 1993), and MLP.

For a fair comparison of classification performance between our proposed framework and other multi-modal fusion methods, we use the same CNN backbones and the MLP classifier as in our proposed architecture. Such comparison avoids the bias caused by differences in backbones and the final classifier.

3.2.4. Data-leakage safeguards

To ensure the consistency of principal components between the training and test sets and prevent data leakage issues, the principal components were learned solely from the training set and then applied to the test set. All atlas-wise operations and scaling were performed at the single sample level, ensuring that no population-level information was leaked to the test set.

3.2.5. Results and discussion

The detailed experimental parameters and results are shown in Table 2. Specifically, we have the following conclusions:

- **Comparing CAMF with linear classifiers:** Comparing Exp 1, 2, and 6, we can see that our proposed framework significantly outperforms all three baseline linear methods according to all evaluation metrics, especially for MCC. This shows the advantage of deep neural networks over linear models.
- **Comparing MLP with linear classifiers:** The comparison between Exp 1–3 shows that MLP does not work well with PCA as the backbone to extract latent features. A reason is that feature selection via PCA and classification with MLP are conducted as distinct, separate processes.
- **Comparing CNNs with PCA:** The comparison between Exp 3 and 4 shows that PCA's effectiveness is compromised by the imbalanced distribution of samples across two classes and PCA cannot capture the essential patterns in the data compared to CNNs as backbones.

Table 2
Binary classification result.

Dataset	Exp ID	Feature extractor	Fusion method	Classifier	F1-Score	Acc	MCC-Score
COBRE +FBIRN +MPRC	1	PCA	Element-wise Sum	SVM	0.6355 ± 0.0377	0.7048 ± 0.0333	0.3968 ± 0.0630
	2	PCA	Element-wise Sum	SGDClassifier	0.6336 ± 0.0453	0.7065 ± 0.0309	0.3979 ± 0.0611
	3	PCA	Element-wise Sum	MLP	0.4529 ± 0.2279	0.4952 ± 0.0613	0.0450 ± 0.1228
	4	CNNs	Element-wise Sum	MLP	0.6684 ± 0.0530	0.7306 ± 0.0352	0.4480 ± 0.0663
	5	CNNs	Concatenation	MLP	0.6506 ± 0.0539	0.7129 ± 0.0273	0.4298 ± 0.0367
	6	CNNs	CAMF	MLP	0.6819 ± 0.0552	0.7339 ± 0.0552	0.4840 ± 0.0876
BSNIP	7	CNNs	Element-wise Sum	MLP	0.6247 ± 0.0366	0.6382 ± 0.0084	0.2831 ± 0.0173
	8	CNNs	Concatenation	MLP	0.6183 ± 0.0555	0.6202 ± 0.0429	0.2416 ± 0.0848
	9	CNNs	CAMF	MLP	0.6656 ± 0.0624	0.6854 ± 0.0362	0.3765 ± 0.0663

- **Comparing CAMF with simple data fusion methods:** The results of Exp 4–6 indicate that the fusion of latent features from multiple modalities may lower the performance if not paired with an appropriate data fusion. Simple concatenation and element-wise sum cannot optimally combine the information from both modalities. The superior performance of CAMF, compared to other fusion methods, demonstrates its ability to extract and merge both inter-modal and intra-modal interactions, resulting in significant improvement. The independent experiments on the BSNIP dataset (Exp 7–9) further corroborate these findings, thereby reinforcing the validity of our conclusions.

3.3. Ablation study

To evaluate the impact of each component in our proposed framework, we conducted an ablation study. Specifically, we assessed the importance of each component by observing the changes in performance resulting from its removal. Here we use the CNN as feature extractors for all scenarios and examine the impact of the two-level data fusion (attention modules and adaptive weights). The results of all scenarios are shown in Table 3, leading to the following conclusions:

- **Uni-modal input for fMRI (Exp 1, 10) and sMRI (Exp 2, 11):** The proposed CAMF framework significantly exceeds the uni-modal methods with respect to all evaluation metrics. This indicates that CAMF can effectively combine the complementary information of multiple modalities to improve classification performance.
- **Simple fusion methods without attention modules (Exp 3–5, 12–14):** Similar to the baselines, the better performance of CAMF displays the advantage of the two-level cross-attentive fusion method proposed in our CAMF framework. Compared to only using the adaptive weights (Exp 5, 14) for the fusion, the results of CAMF underscore the enhanced predictive capability brought by the SA and CA modules.
- **Removing cross-attention or self-attention modules (Exp 5–7, 14–16):** The experimental results show that the cross-attention module (Exp 7, 16) can significantly improve the performance compared to not using it (Exp 5, 14). This indicates that the CA modules could well exploit the interaction between modalities, thereby leading to better classification results. Although employing only the SA modules (Exp 6, 15) cannot independently improve the performance, their combination with the CA modules proves to be beneficial. This proves that both CA and SA are indispensable components of the proposed method.
- **Removing adaptive weights (Exp 8, 17):** Finally, we compare the fusion method with adaptive weights with simple concatenation (Exp 8, 17). This shows that the fusion with optimal weight can effectively incorporate the information from inter- and within-modality interactions.

We further performed Leave-One-Site-Out (LOSO-CV) to explore the variations between sites. The results in the supplementary materials show that CAMF captures robust features across different sites and achieves the best performance while combining multiple datasets.

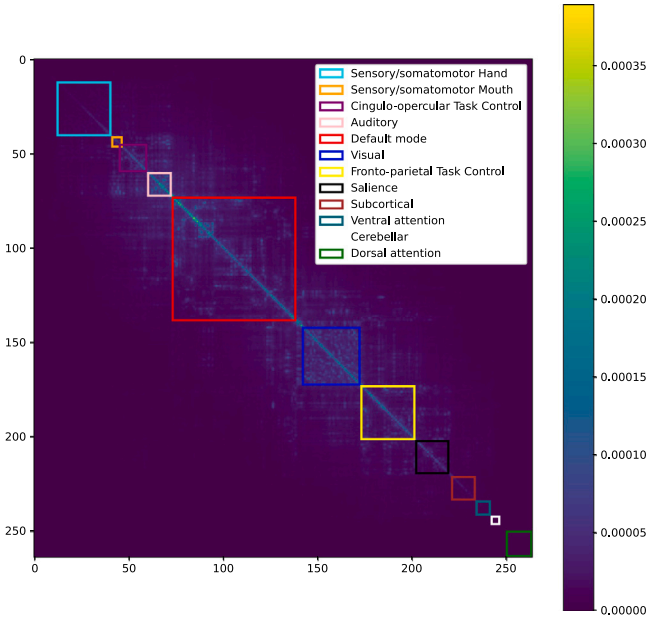

Fig. 3. Average saliency map from the FC modality, where the rows and columns in the FC matrix are grouped by different functional brain networks, which are highlighted in various colored boxes. We observe a high correlation with each box.

The ablation study showed the importance of each modality and underscored the contributions of each component within the proposed framework. However, it cannot give the interpretation of the model behaviors at the region level or even pixel/voxel level, which is more comprehensible. Thus, we further utilized the Score-CAM method (Wang et al., 2020) to generate high-resolution saliency maps that help improve the understanding of the model and identify biomarkers associated with schizophrenia.

3.4. Interpretation

When predicting schizophrenia identification, the regions highlighted by gradient-guided Score-CAM can be interpreted as the key disease-related brain functional networks (BFN) and brain structural regions (BSR).

Using the Score-CAM method at the last convolutional layer of each backbone, we produced two saliency maps for each subject with low resolutions (6×6 for fMRI and $5 \times 7 \times 5$ for sMRI) due to the pooling layers in the backbones. Then we interpolated the saliency maps to match with the size of the input data (264×264 for fMRI and $121 \times 145 \times 121$ for sMRI). The high-resolution saliency maps after interpolation allow biomarker identification from the input data. In order to identify the disease-related BFNs and BSRs among the whole

Table 3
The prediction results (mean and std) of the ablation study.

Dataset	Exp ID	Modality type	Fusion method	F1-Score	Acc	MCC-Score	Class-weighted Loss	AUCROC	Ave PRC
COBRE +FBIRN +MPRC	1	Uni-modal (fMRI)	–	0.5601 ± 0.0640	0.6645 ± 0.0181	0.2989 ± 0.0627	1.2922 ± 0.0405	0.6973 ± 0.0215	0.6394 ± 0.0585
	2	Uni-modal (sMRI)	–	0.6392 ± 0.0518	0.6952 ± 0.0236	0.3959 ± 0.0466	4.3698 ± 0.8946	0.7525 ± 0.0229	0.6978 ± 0.0829
	3	Multi-modal	Element-wise Sum	0.6684 ± 0.0530	0.7306 ± 0.0352	0.4480 ± 0.0663	3.0423 ± 0.4182	0.7900 ± 0.0179	0.7527 ± 0.0704
	4	Multi-modal	Concatenation	0.6506 ± 0.0539	0.7129 ± 0.0273	0.4298 ± 0.0367	3.7253 ± 1.8587	0.7879 ± 0.0186	0.7358 ± 0.0760
	5	Multi-modal	Adaptive Weights	0.6133 ± 0.0545	0.6952 ± 0.0172	0.3848 ± 0.0434	2.2433 ± 0.2854	0.7713 ± 0.0156	0.7467 ± 0.0469
	6	Multi-modal	2 Self-attention + Adaptive Weights	0.6375 ± 0.0520	0.7145 ± 0.0397	0.4258 ± 0.0683	1.8324 ± 0.4163	0.7841 ± 0.0198	0.7519 ± 0.0603
	7	Multi-modal	2 Cross-attention + Adaptive Weights	0.6592 ± 0.0683	0.7274 ± 0.0351	0.4484 ± 0.0743	**1.2846 ± 0.0751**	0.8017 ± 0.0197	0.7778 ± 0.0523
	8	Multi-modal	2 Cross-attention + 2 Self-attention + Concatenation	0.6480 ± 0.0595	0.7032 ± 0.0369	0.3929 ± 0.0733	1.9602 ± 0.3679	0.7609 ± 0.0475	0.7255 ± 0.0885
	9	Multi-modal	CAMF	**0.6819 ± 0.0552**	**0.7339 ± 0.0552**	**0.4840 ± 0.0876**	1.5829 ± 0.2512	**0.8074 ± 0.0410**	**0.7929 ± 0.0660**
BSNIP	10	Uni-modal (fMRI)	–	0.4647 ± 0.0386	0.5753 ± 0.0149	0.1748 ± 0.0402	**1.3784 ± 0.0064**	0.5804 ± 0.0090	0.6629 ± 0.0096
	11	Uni-modal (sMRI)	–	0.6002 ± 0.0473	0.6360 ± 0.0231	0.2847 ± 0.0503	3.5711 ± 0.2497	0.7054 ± 0.0208	0.7090 ± 0.0134
	12	Multi-modal	Element-wise Sum	0.6247 ± 0.0366	0.6382 ± 0.0084	0.2831 ± 0.0173	3.2770 ± 0.6303	0.7153 ± 0.0292	0.7134 ± 0.0183
	13	Multi-modal	Concatenation	0.6183 ± 0.0555	0.6202 ± 0.0429	0.2416 ± 0.0848	3.9343 ± 0.4891	0.6881 ± 0.0409	0.6939 ± 0.0494
	14	Multi-modal	Adaptive Weights	0.6220 ± 0.0594	0.6449 ± 0.0298	0.2979 ± 0.0590	2.5878 ± 0.2700	0.7010 ± 0.0204	0.6937 ± 0.0245
	15	Multi-modal	2 Self-attention + Adaptive Weights	0.6252 ± 0.0541	0.6382 ± 0.0321	0.2789 ± 0.0614	1.7746 ± 0.4558	0.7029 ± 0.0149	0.7090 ± 0.0407
	16	Multi-modal	2 Cross-attention + Adaptive Weights	0.6060 ± 0.0575	0.6562 ± 0.0242	0.3288 ± 0.0410	1.7536 ± 0.3818	0.7196 ± 0.0184	0.7186 ± 0.0135
	17	Multi-modal	2 Cross-attention + 2 Self-attention + Concatenation	0.5928 ± 0.0746	0.6337 ± 0.0272	0.2765 ± 0.0433	1.9642 ± 0.4071	0.6773 ± 0.0368	0.6939 ± 0.0323
	18	Multi-modal	CAMF	**0.6656 ± 0.0624**	**0.6854 ± 0.0362**	**0.3765 ± 0.0663**	1.8004 ± 0.3329	**0.7453 ± 0.0164**	**0.7387 ± 0.0263**

population, for each modality, we generated an average template by averaging the saliency maps of all subjects. Like the gradient-guided CAM method (Selvaraju et al., 2017), we also applied element-wise multiplication of gradients of the predicted class to the saliency maps to generate the refined regions.

For the fMRI, the input FC is always symmetric since it is derived by Pearson correlation. However, the saliency map could be asymmetric due to the padding of kernels in CNNs. Thus we used the average of FC saliency map template and its transpose to address the asymmetry issue. The results are shown in Fig. 3.

For the sMRI, we first visualize the voxel-level average saliency map template from the x-, y- and *z*-axis in Fig. 4(a)–(c). In order to locate the disease-related BSRs, we then segment the template using the automated anatomical labeling (AAL) (Tzourio-Mazoyer et al., 2002) atlas. The voxels in each anatomical BSR are grouped together and the BSRs are ranked by their averaged activation scores. The region-level saliency map template is shown in Fig. 4(d)–(f).

4. Discussion

4.1. Schizophrenia-related brain region identification

The saliency map template from the fMRI data shows the key disease-related pair-wise functional connectivities. According to Fig. 3, most highlighted BFNs lie in the auditory, default mode, and visual networks. The high activation scores indicate the connections within these regions to schizophrenia, which aligns with the results from previous research. The auditory network is considered responsible for receiving and processing sound, which is the basis of comprehension and analysis of the meaning of utterances (Hackett, 2015). For example, a review study (Sun et al., 2009) reported a significant difference in the superior temporal gyrus (in the auditory network) or subregional volume, including bilateral or unilateral ROI, and that volume reduction was the common change in patients with schizophrenia. The visual network is involved with processing visual information, and another research (Ford et al., 2015) presented evidence that SZs who endorse having recently experienced visual hallucinations have hyperconnectivity between the amygdala and visual cortex. The default mode network is a set of brain regions that become more active when the individual is in a resting state (Buckner, 2012). A study (Garrity et al., 2007) concluded that in the default mode network, both temporal frequency alterations and disruption of local spatial patterns are associated with schizophrenia.

The saliency map template of sMRI in Fig. 4(a)–(c) shows the voxel-level intensity in the 3D space that the model utilizes for the classification. By ranking the BSRs according to their average activation scores, the top three highlighted BSRs are cingulum, thalamus, and caudate. The abnormalities of cingulum bundle, one of the most distinctive fiber tracts in the brain, have been linked to schizophrenia by several prior studies (Fitzsimmons et al., 2020; Whitford et al., 2014). The thalamus is primarily a gray matter structure within the diencephalon, which plays a key role in linking and relaying information between brain regions, including the visual system and the primary auditory cortex. This matches the disease-related BFNs identified by the FC. Prior research also found the thalamus to be a region associated with schizophrenia (Pergola et al., 2015). The study in Dorph-Petersen and Lewis (2017) revealed a correlation between a reduction in volume and cell numbers in the pulvinar and the most robust evidence for structural changes of the thalamus in schizophrenia. The caudate nucleus plays a critical role in several higher cognitive functions, such as memory, language, and emotion. A previous study (Takase et al., 2004) identified that the caudate is associated with schizophrenia, since the caudate nucleus of patients has smaller volumes compared to HCs. Another study (Jiang et al., 2023) characterized two distinct but stable 'trajectories' of brain atrophy, separately beginning in the Broca's area (subtype1) and the hippocampus (subtype2). Future studies are needed to decode the correlation between the BSRs identified by CAMF and the distinct pathophysiological processes underlying schizophrenia, which is the key to potential applications such as prognosis, personalized treatment, and monitoring.

We further performed Score-CAM ablation by restricting the input to a single modality (either fMRI or sMRI) while maintaining the self-attention and dynamic weight modules. We also performed a sanity check to validate the feasibility of our interpretation method and the importance of the highlighted voxels in the saliency maps. The results are included in the supplementary materials.

4.2. Sex impact on Schizophrenia-related brain structural regions

We further investigate the impact of gender on the voxel-level sMRI saliency map. We firstly select the SZ patients of the major age range (20–60 years old). And then generate the averaged saliency map for the males and females separately, as shown in 5(a)–(b). Additionally, we also calculate voxel-level difference of the averaged saliency map between the two gender groups in 5(c). The saliency maps show little voxel-level difference between males and females. And the averaged

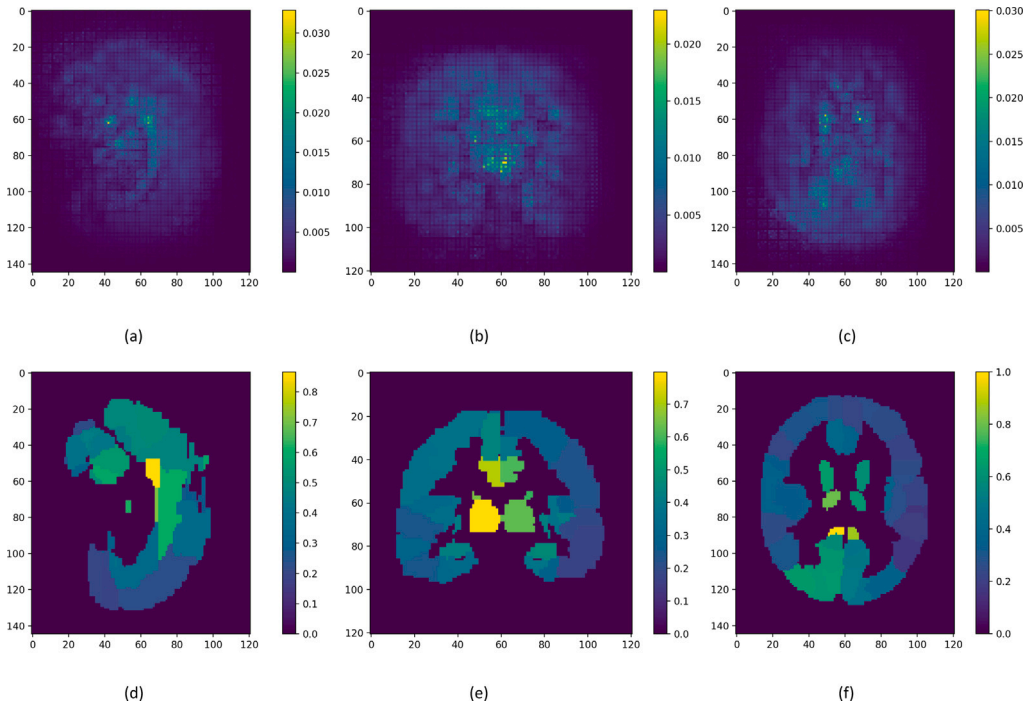

Fig. 4. Voxel-level and region-level Score-CAM saliency maps from x-, y- and z- axis. In the region-level saliency maps, the intensity is averaged inside each brain region defined by the automated anatomical labeling (AAL) (Tzourio-Mazoyer et al., 2002) atlas. Subfigures (a)–(c) show the voxel-level average saliency maps and subfigures (d)–(f) show the region-level average saliency maps. For the subfigure of each axis, we choose the slice along the medial axis to visualize the highlighted voxels/regions at the center of the brain.

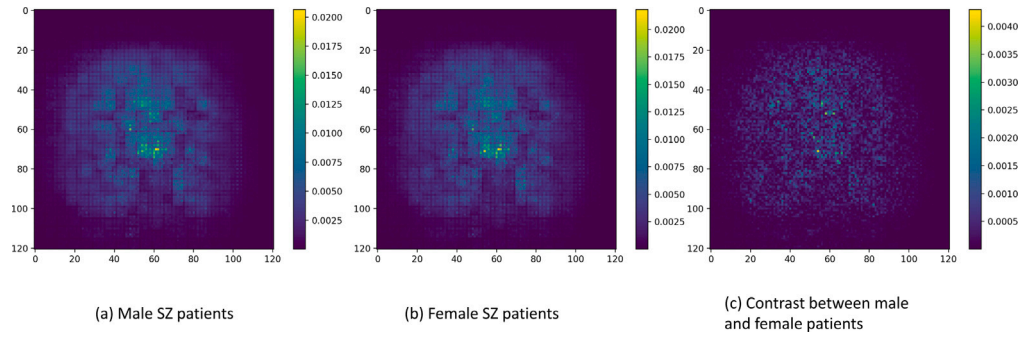

Fig. 5. Voxel-level Score-CAM saliency maps of 20–60 year old schizophrenia patients from y-axis. (a) shows the voxel-level average saliency maps of male SZ patients, (b) is for female patients and (c) is the contrast of saliency map between males and females.

saliency maps of both genders align with that of all samples, proving the stability of our method.

The established relationships between BFNs (auditory, default mode, and visual networks) identified from fMRI and BSRs (cingulum, thalamus, and caudate) identified from sMRI validated the reliability and interpretability of our proposed framework. The similar findings of previous studies also indicate the potential clinical application of our data-driven framework.

We also reported the stratified performance results for sex and age in the supplementary materials to evaluate the impact of confounding factors such as age and sex. Our previous paper (Orlichenko et al., 2024) developed a variational autoencoder-based framework to mitigate demographic confounds.

4.3. Limitations and future scope

Although our proposed framework successfully improved the classification performance by fusing multi-modal data effectively, there are still limitations and thus potential avenues for future research. First, our framework deployed a 2D CNN as the backbone to extract latent features based on the FCs. Employing a graph-based representation for FCs or using a graph neural network as we did in Qu et al. (2021a,b), Wang et al. (2021) may better capture structural information, potentially elevating both performance and interpretability. We plan to integrate GNN backbones in the future framework. Second, although Score-CAM can produce saliency maps on CNN backbones, other interpretable methods can be exploited for potential improvements. Third, while our proposed framework is tested for the two modalities, it can be generalized to three or more modalities (e.g., SNPs), which may further improve the prediction performance. In addition, although it is still an open problem with model architecture design on the order of SA/CA modules, combining CA and SA modules sequentially could be an beneficial alternative compared to the parallel setting used in CAMF. Further experiments are needed to investigate its impact on performance. Last, the way in constructing FC may affect the model performance. Further testing of using various parcellation atlases (e.g., Schaefer-400 and Glasser) and correlation methods (e.g., Fisher-z or partial correlation) could help validate the robustness of CAMF.

5. Conclusion

We proposed a novel deep learning framework for fusing heterogeneous fMRI and sMRI data, leading to improved classification accuracy in identifying schizophrenia relative to prior models. Building on transformer architectures, we incorporated self-attention and cross-attention modules to incorporate the intra- and inter-modality relationships to better represent multi-modal brain imaging data. The experimental results on multiple cohorts demonstrated the superiority of our method over conventional multi-modal fusion methods. In addition, Score-CAM was used to identify critical disease-related functional networks (the auditory, default mode, and visual networks) and structural regions (cingulum, thalamus, and caudate). These findings are in concordance with the results reported in prior studies, validating the framework in biomarker discovery and disease identification.

CRediT authorship contribution statement

Ziyu Zhou: Writing – original draft, Software, Methodology. **Anton Orlichenko:** Writing – review & editing, Methodology. **Gang Qu:** Writing – review & editing, Methodology, Conceptualization. **Zening Fu:** Writing – review & editing, Data curation. **Zhengming Ding:** Writing – review & editing, Methodology. **Julia Stephen:** Writing – review & editing, Investigation, Conceptualization. **Tony Wilson:** Writing – review & editing, Investigation. **Vince Calhoun:** Writing – review & editing. **Yu-Ping Wang:** Writing – review & editing, Supervision, Investigation, Funding acquisition.

Ethics statements

The datasets used were acquired from existing public data repositories. This study does not include new human-subject experiments. All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional review board (IRB).

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 in part by National Institutes of Health (NIH) under grants R01 GM109068, R01 MH104680, R01 MH107354, P20 GM103472, R01 EB020407, R01EB006841, R01EB036247, R56MH124925, R01MH121101, P20GM144641, and 2U54MD007595, and in part by National Science Foundation under grants 1539067 and 2112455.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ynirp.2026.100338.

Data availability

- The BSNIP dataset is available online at https://nda.nih.gov/edit_collection.html?id=2274.
- The COBRE dataset is available by using the COINS platform at https://coins.trendscenter.org/.
- The FBIRN dataset is available online at https://www.nitrc.org/projects/fbirn/.
- The MPRC dataset can be requested from the corresponding authors of the original article.
- The code of our proposed method is available upon request.

References

Adhikari, B.M., et al., 2019. Functional network connectivity impairments and core cognitive deficits in schizophrenia. Hum. Brain Mapp. 40 (16), 4593–4605.
Aine, C., et al., 2017. Multimodal neuroimaging in schizophrenia: description and dissemination. Neuroinformatics 15, 343–364.
Amari, S.-i., 1993. Backpropagation and stochastic gradient descent method. Neurocomputing 5 (4–5), 185–196.
Buckner, R.L., 2012. The serendipitous discovery of the brain's default network. Neuroimage 62 (2), 1137–1145.
Calhoun, V.D., Sui, J., 2016. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. Biological Psychiatry: Cogn. Neurosci. Neuroimaging 1 (3), 230–244.
Chen, L., Qiao, C., Ren, K., Qu, G., Calhoun, V.D., Stephen, J.M., Wilson, T.W., Wang, Y.-P., 2024. Explainable spatio-temporal graph evolution learning with applications to dynamic brain network analysis during development. NeuroImage 298, 120771.
Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21 (1), 1–13.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Dorph-Petersen, K.-A., Lewis, D.A., 2017. Postmortem structural studies of the thalamus in schizophrenia. Schizophr. Res. 180, 28–35.
Fan, H., Liu, Z., Wu, X., Yu, G., Gu, X., Kuang, N., Zhang, K., Liu, Y., Jia, T., Sahakian, B.J., et al., 2023. Decoding anxiety–impulsivity subtypes in preadolescent internalising disorders: findings from the adolescent brain cognitive development study. Br. J. Psychiatry 223 (6), 542–554.
Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neurosci. 18 (11), 1664–1671.
Fitzsimmons, J., et al., 2020. Cingulum bundle abnormalities and risk for schizophrenia. Schizophr. Res. 215, 385–391.
Ford, J.M., Palzes, V.A., Roach, B.J., Potkin, S.G., van Erp, T.G., Turner, J.A., Mueller, B.A., Calhoun, V.D., Voyvodic, J., Belger, A., et al., 2015. Visual hallucinations are associated with hyperconnectivity between the amygdala and visual cortex in people with a diagnosis of schizophrenia. Schizophr. Bull. 41 (1), 223–232.
Gao, Y., Lewis, N., Calhoun, V.D., Miller, R.L., 2023. Interpretable LSTM model reveals transiently-realized patterns of dynamic brain connectivity that predict patient deterioration or recovery from very mild cognitive impairment. Comput. Biol. Med. 161, 107005.
Garrity, A.G., Pearlson, G.D., McKiernan, K., Lloyd, D., Kiehl, K.A., Calhoun, V.D., 2007. Aberrant "default mode" functional connectivity in schizophrenia. Am. J. Psychiatry 164 (3), 450–457.
Hackett, T.A., 2015. Anatomic organization of the auditory cortex. Handb. Clin. Neurol. 129, 27–53.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.
Hojjati, S.H., Ebrahimzadeh, A., Babajani-Feremi, A., 2019. Identification of the early stage of alzheimer's disease using structural MRI and resting-state fMRI. Front. Neurol. 10, 904.
Jiang, Y., Wang, J., Zhou, E., Palaniyappan, L., Luo, C., Ji, G., Yang, J., Wang, Y., Zhang, Y., Huang, C.-C., et al., 2023. Neuroimaging biomarkers define neurophysiological subtypes with distinct trajectories in schizophrenia. Nat. Ment. Health 1 (3), 186–199.
Katuwal, G.J., Cahill, N.D., Baum, S.A., Michael, A.M., 2015. The predictive power of structural MRI in autism diagnosis. In: 2015 37th Annual International Conference of the Ieee Engineering in Medicine and Biology Society. EMBC, pp. 4270–4273.
Keator, D.B., et al., 2016. The function biomedical informatics research network data repository. Neuroimage 124, 1074–1079.
Khalilullah, K.I., Agcaoglu, O., Sui, J., Adali, T., Duda, M., Calhoun, V.D., 2023. Multimodal fusion of multiple rest fMRI networks and MRI gray matter via parallel multilink joint ICA reveals highly significant function/structure coupling in alzheimer's disease. Hum. Brain Mapp. 44 (15), 5167–5179.
Li, H., Satterthwaite, T.D., Fan, Y., 2018. Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks. In: 2018 Ieee 15th International Symposium on Biomedical Imaging (Isbi 2018). pp. 101–104.
Liu, Y., Peng, S., Wu, X., Liu, Z., Lian, Z., Fan, H., Kuang, N., Gu, X., Yang, S., Hu, Y., et al., 2025. Neural, cognitive and psychopathological signatures of a prosocial or delinquent peer environment during early adolescence. Dev. Cogn. Neurosci. 73, 101566.
Liu, L., Wang, Y.-P., Wang, Y., Zhang, P., Xiong, S., 2022. An enhanced multi-modal brain graph network for classifying neuropsychiatric disorders. Med. Image Anal. 81, 102550.
Meng, X., Iraji, A., Fu, Z., Kochunov, P., Belger, A., Ford, J.M., McEwen, S., Mathalon, D.H., Mueller, B.A., Pearlson, G., et al., 2023. Multi-model order spatially constrained ICA reveals highly replicable group differences and consistent predictive results from resting data: A large n fMRI schizophrenia study. NeuroImage: Clin. 38, 103434.

Miller, B., et al., 2011. Meta-analysis of paternal age and schizophrenia risk in male versus female offspring. Schizophr. Bull. 37 (5), 1039–1047.
Mousavian, M., Chen, J., Traylor, Z., Greening, S., 2021. Depression detection from sMRI and rs-fMRI images using machine learning. J. Intell. Inf. Syst. 57, 395–418.
Oh, J., Oh, B.-L., Lee, K.-U., Chae, J.-H., Yun, K., 2020. Identifying schizophrenia using structural MRI with a deep learning algorithm. Front. Psychiatry 11, 16.
Orlichenko, A., Qu, G., Zhou, Z., Liu, A., Deng, H.-W., Ding, Z., Stephen, J.M., Wilson, T.W., Calhoun, V.D., Wang, Y.-P., 2024. A demographic-conditioned variational autoencoder for fmri distribution sampling and removal of confounds. ArXiv.arXiv–2405.
Pergola, G., Selvaggi, P., Trizio, S., Bertolino, A., Blasi, G., 2015. The role of the thalamus in schizophrenia from a neuroimaging perspective. Neurosci. Biobehav. Rev. 54, 57–75.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. Neuroimage 59 (3), 2142–2154.
Power, J.D., et al., 2011. Functional network organization of the human brain. Neuron 72 (4), 665–678.
Qu, G., Hu, W., Xiao, L., Wang, J., Bai, Y., Patel, B., Zhang, K., Wang, Y.-P., 2021a. Brain functional connectivity analysis via graphical deep learning. IEEE Trans. Biomed. Eng. 69 (5), 1696–1706.
Qu, G., Orlichenko, A., Wang, J., Zhang, G., Xiao, L., Zhang, K., Wilson, T.W., Stephen, J.M., Calhoun, V.D., Wang, Y.-P., 2023. Interpretable cognitive ability prediction: A comprehensive gated graph transformer framework for analyzing functional brain networks. IEEE Trans. Med. Imaging.
Qu, G., et al., 2021b. Ensemble manifold regularized multi-modal graph convolutional network for cognitive ability prediction. IEEE Trans. Biomed. Eng. 68 (12), 3564–3573.
Rahaman, M.A., Chen, J., Fu, Z., Lewis, N., Iraji, A., van Erp, T.G., Calhoun, V.D., 2023. Deep multimodal predictome for studying mental disorders. Hum. Brain Mapp. 44 (2), 509–522.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.
Sui, J., et al., 2013. Three-way (n-way) fusion of brain imaging data based on mCCA+ jICA and its application to discriminating schizophrenia. NeuroImage 66, 119–132.
Sun, J., Maller, J.J., Guo, L., Fitzgerald, P.B., 2009. Superior temporal gyrus volume change in schizophrenia: a review on region of interest volumetric studies. Brain Res. Rev. 61 (1), 14–32.
Takase, K., Tamagaki, C., Okugawa, G., Nobuhara, K., Minami, T., Sugimoto, T., Sawada, S., Kinoshita, T., 2004. Reduced white matter volume of the caudate nucleus in patients with schizophrenia. Neuropsychobiology 50 (4), 296–300.
Tamminga, C.A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., Thaker, G., 2014. Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. Schizophr. Bull. 40 (Suppl_2), S131–S137.
Tomassini, S., Falcionelli, N., Sernani, P., Müller, H., Dragoni, A.F., 2021. An end-to-end 3D convlstm-based framework for early diagnosis of alzheimer's disease from full-resolution whole-brain sMRI scans. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems. CBMS, pp. 74–78.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI mri single-subject brain. Neuroimage 15 (1), 273–289.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X., 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 24–25.
Wang, J., et al., 2021. Functional network estimation using multigraph learning with application to brain maturation study. Hum. Brain Mapp. 42 (9), 2880–2892.
Whitford, T.J., et al., 2014. Localized abnormalities in the cingulum bundle in patients with schizophrenia: a diffusion tensor tractography study. NeuroImage: Clin. 5, 93–99.
Wu, X., Zhang, K., Kuang, N., Kong, X., Cao, M., Lian, Z., Liu, Y., Fan, H., Yu, G., Liu, Z., et al., 2025. Developing brain asymmetry shapes cognitive and psychiatric outcomes in adolescence. Nat. Commun. 16 (1), 4480.
Yan, W., Pearlson, G.D., Fu, Z., Li, X., Iraji, A., Chen, J., Sui, J., Volkow, N.D., Calhoun, V.D., 2024. A brainwide risk score for psychiatric disorder evaluated in a large adolescent population reveals increased divergence among higher-risk groups relative to control participants. Biol. Psychiatry 95 (7), 699–708.
Yang, Y., Ye, C., Guo, X., Wu, T., Xiang, Y., Ma, T., 2023. Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. IEEE Trans. Med. Imaging.
Zhang, Y., Zhang, H., Chen, X., Lee, S.-W., Shen, D., 2017. Hybrid high-order functional connectivity networks using resting-state functional MRI for mild cognitive impairment diagnosis. Sci. Rep. 7 (1), 6530.
Zhu, Q., Wang, H., Xu, B., Zhang, Z., Shao, W., Zhang, D., 2022. Multimodal triplet attention network for brain disease diagnosis. IEEE Trans. Med. Imaging 41 (12), 3884–3894.
Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., Initiative, A.D.N., 2016. Label-aligned multi-task feature learning for multimodal classification of alzheimer's disease and mild cognitive impairment. Brain Imaging Behav. 10, 1148–1159.