Psychoradiology, 2024, **4**, kkae005

DOI: 10.1093/psyrad/kkae005

Advance access publication date: 26 March 2024

Research Article

Transcriptomic and neuroimaging data integration enhances machine learning classification of schizophrenia

Mengya Wang[1], Shu-Wan Zhao[1,2], Di Wu[3], Ya-Hong Zhang[4], Yan-Kun Han[2], Kun Zhao[1], Ting Qi[5], Yong Liu[1], Long-Biao Cui[2,6,7,8,*] and Yongbin Wei[1,*]

[1], Center for Artificial Intelligence in Medical Imaging, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]Schizophrenia Imaging Lab, Xijing 986 Hospital, Fourth Military Medical University, Xi'an, 710054, China
[3]Department of Psychiatry, Xijing Hospital, Fourth Military Medical University, Xi'an, 710032, China
[4]Department of Psychiatry, Xi'an Gaoxin Hospital, Xi'an, 710075, China
[5]Department of Neurology, School of Medicine, University of California San Francisco, San Francisco, 94143, California
[6]Department of Radiology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061, China
[7]Department of Radiology, The Second Medical Center, Chinese PLA General Hospital, Beijing, 100853, China
[8]Shaanxi Provincial Key Laboratory of Clinic Genetics, Fourth Military Medical University, Xi'an, 710032, China
*Correspondence: Yongbin Wei, yongbin.wei@bupt.edu.cn; Long-biao Cui, lbcui@fmmu.edu.cn

Abstract

Background: Schizophrenia is a polygenic disorder associated with changes in brain structure and function. Integrating macroscale brain features with microscale genetic data may provide a more complete overview of the disease etiology and may serve as potential diagnostic markers for schizophrenia.

Objective: We aim to systematically evaluate the impact of multi-scale neuroimaging and transcriptomic data fusion in schizophrenia classification models.

Methods: We collected brain imaging data and blood RNA sequencing data from 43 patients with schizophrenia and 60 age- and gender-matched healthy controls, and we extracted multi-omics features of macroscale brain morphology, brain structural and functional connectivity, and gene transcription of schizophrenia risk genes. Multi-scale data fusion was performed using a machine learning integration framework, together with several conventional machine learning methods and neural networks for patient classification.

Results: We found that multi-omics data fusion in conventional machine learning models achieved the highest accuracy (AUC ∼0.76–0.92) in contrast to the single-modality models, with AUC improvements of 8.88 to 22.64%. Similar findings were observed for the neural network, showing an increase of 16.57% for the multimodal classification model (accuracy 71.43%) compared to the single-modal average. In addition, we identified several brain regions in the left posterior cingulate and right frontal pole that made a major contribution to disease classification.

Conclusion: We provide empirical evidence for the increased accuracy achieved by imaging genetic data integration in schizophrenia classification. Multi-scale data fusion holds promise for enhancing diagnostic precision, facilitating early detection and personalizing treatment regimens in schizophrenia.

Keywords: schizophrenia; machine learning; multi-omics; genomics; transcriptomics

Introduction

Schizophrenia (SZ) is a prevalent neuropsychiatric condition characterized by symptoms of hallucinations, delusions, cognitive deficits, and emotional disturbances (McCutcheon *et al.*, 2020). Using multimodal neuroimaging data, patients with SZ have been observed to display a wide range of abnormalities in brain morphology (Hulshoff Pol *et al.*, 2002; Liu *et al.*, 2020) as well as the structural and functional connectome (Cui *et al.*, 2019; Griffa *et al.*, 2019; Gao *et al.*, 2023). Given such observed brain abnormalities, emerging research implements conventional machine learning (ML) or deep learning (DL) frameworks to distinguish SZ patients from healthy individuals, aiming to build up an objective, valid model that augments diagnosis or prognosis of the disorder in clinical practice (Gao *et al.*, 2018; Sadeghi *et al.*, 2022; Chen *et al.*, 2023; Sui *et al.*, 2023). Despite these efforts, the extant methods have yet to achieve clinical applicability. Possible reasons that hinder clinical translation include the high heterogeneity among patients and the highly complex etiology across multiple scales (Guggenmos *et al.*, 2020; Sadeghi *et al.*, 2022).

Integrating neuroimaging data with genetic data might improve our understanding of the complex heterogeneity in SZ pathophysiology and further enhance the performance of ML/DL models. SZ is known to be a polygenic disorder determined by multiple genetic variants (Trubetskoy *et al.*, 2022). More intriguingly, SZ polygenic risk scores have been observed to be associated with the macroscale connectomic changes in the brain (Cao

Received: 4 December 2023; **Revised:** 26 February 2024; **Accepted:** 25 March 2024

et al., 2020; Wei *et al.*, 2023), suggesting that connectivity deficits in SZ patients might be related to specific genetic variants. Furthermore, transcriptional profiles of SZ risk genes were found to be associated with brain volume changes and disconnectivity profiles in SZ (Romme *et al.*, 2017; Ji *et al.*, 2021), pointing to the multi-scale association of SZ etiology. These results also implicated the potential of fusing multi-scale, multi-omics data to explain the disorder more accurately. A study has indeed found an increased explained variance when classifying SZ from healthy controls (HCs) using brain characteristics to fine-tune polygenic scores of SZ (van der Meer *et al.*, 2022). However, it remains unknown whether combining transcriptomic and neuroimaging data could improve ML/DL models for SZ classification, and if so, how different types of data fusion method behave.

To this end, the current study collected blood transcriptomic and neuroimaging data and extracted multi-omics features to evaluate the impact of multi-scale data fusion in SZ classification. Three ML integration frameworks were evaluated, together with several conventional ML methods and DL neural networks. We also employed systematic ablation experiments to assess the contributions of each dataset, which enhanced the interpretability of the models. We hypothesize that integrating transcription data with neuroimaging will increase classification accuracy in different settings of SZ classifiers.

Materials and methods

Participants

The current study includes 43 patients with SZ and 60 age- and sex-matched HC. All participants were right-handed and of Han Chinese ethnicity. Patients were recruited from the Department of Psychiatry at Xijing Hospital and controls were enrolled from local communities through advertising. Patients were diagnosed based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5), with consensus diagnoses made by two experienced clinical psychiatrists using all the available information. This study was approved by the institutional ethics committee, First Affiliated Hospital of Fourth Military Medical University and all participants provided written informed consent of participation.

RNA-seq acquisition and preprocessing

RNA sequencing (RNA-seq) data from intravenous blood were used in this study. The data collection protocol has been described in detail in our previous study (Cui *et al.*, 2023). Briefly, blood samples (2.5 ml) were collected in a PAXgene Blood RNA Tube and were immediately frozen at −80°C. RNA sequencing was performed using Illumina Novaseq 6000 (Rothberg *et al.*, 2011). Low-quality reads were filtered out by Fastp (v.0.18.0) (Chen *et al.*, 2018) and filtered reads were mapped to human reference genome hg19 using HISAT2.2.4 (Kim *et al.*, 2015). The count data of 20 313 genes representing the number of sequence reads were obtained. Count data were further normalized using DESeq2 (Love *et al.*, 2014), resulting in normalized gene expression of 17 999 genes.

We then selected SZ risk genes using summary statistics from the largest genome-wide association study (GWAS) on SZ from East Asian populations (including 22 778 patients with SZ and 35 362 control participants) (Lam *et al.*, 2019). The SNP-based statistics were mapped to 346 genes using three gene mapping approaches, including positional mapping, eQTL mapping, and chromatin interaction mapping, which were performed using FUMA (Watanabe *et al.*, 2017). A total of 346 genes were then obtained and defined as SZ risk genes, and their corresponding gene expression data were used in the following analysis (Supplementary Table 1).

Table 1: Scanning parameters.

	T1	DWI	rsfMRI
TR (ms)	8.2	10 000	2000
TE (ms)	3.2	82.4	30
Flip angle (°)	12	NA	90
Field of view (mm^2)	256 × 256	240 × 240	240 × 240
Matrix	256 × 256	128 × 128	64 × 64
Slice thickness (mm)	1	2	3.5
Slice gap (mm)	0	0	0
Number of slices	196	70	45

NA, not applicable; TE, echo time; TR, repetition time.

Neuroimaging data acquisition and preprocessing

T1-weighted magnetic resonance imaging (MRI), diffusion-weighted imaging (DWI), and resting-state functional MRI (rsfMRI) data were collected using a GE Discovery MR750 3.0 T scanner. Scanning parameters were described in our previous study (Cui *et al.*, 2019) and are tabulated in Table 1.

T1-weighted MRI data were preprocessed using FreeSurfer (v.6.0) (Fischl, 2012) to segment brain tissue and to reconstruct cortical mantle. Using the Desikan–Killiany (Desikan *et al.*, 2006) atlas, the reconstructed cortical ribbon and subcortex were divided into 82 brain regions [68 cortical regions (34 in each hemisphere) and 14 subcortical regions] (Desikan *et al.*, 2006).

Connectome reconstruction was conducted using DWI data and rsfMRI data through FSL (v.6.0) (Jenkinson *et al.*, 2012) and CATO (v.3.1.2) (de Lange *et al.*, 2021). Briefly, DWI data processing includes: (i) volume realignment and corrections; (ii) diffusion peaks reconstruction via CSD (Morez *et al.*, 2021); and (iii) fiber tract reconstruction using FACT and streamlined tractography (Mori *et al.*, 1999). For each participant, an 82 × 82 structural connectivity (SC) matrix was reconstructed. The streamline density (i.e. the number of streamlines between two regions divided by region volume) was used as SC weight. To exclude false-positive connections (de Reus and van den Heuvel, 2013), connections that present in >60% of the entire sample were selected in the current studies. Missing values (i.e. unknown clinical features or connections that could not be detected in patients) in the sample were supplemented with the mean value of this feature across all participants.

RsfMRI data processing includes: (i) slice timing, realignment and co-registration with T1-weighted image; (ii) linear trends correction of the blood oxygenation level-dependent (BOLD) time series, as well as global nuisance covariance, including head motion parameters and mean signals of white matter and ventricles; and (iii) band-pass filtering and motion scrubbing with FD and DVARS thresholds (FD > 0.25, DVARS > 1.5). The functional connectivity matrix describes correlations of the extracted BOLD time series between every two regions.

Feature extraction

After preprocessing transcriptomic and neuroimaging data, 38 SZ and 48 HC were included (5 SZ and 12 HC were excluded due to missing data). Specifically, 4 SZ and 11 HC were excluded due to missing FC data; 1 SZ and 1 HC were excluded due to the absence of cortical thickness data. For each of the remaining participants, five feature matrices were obtained, including brain volume,

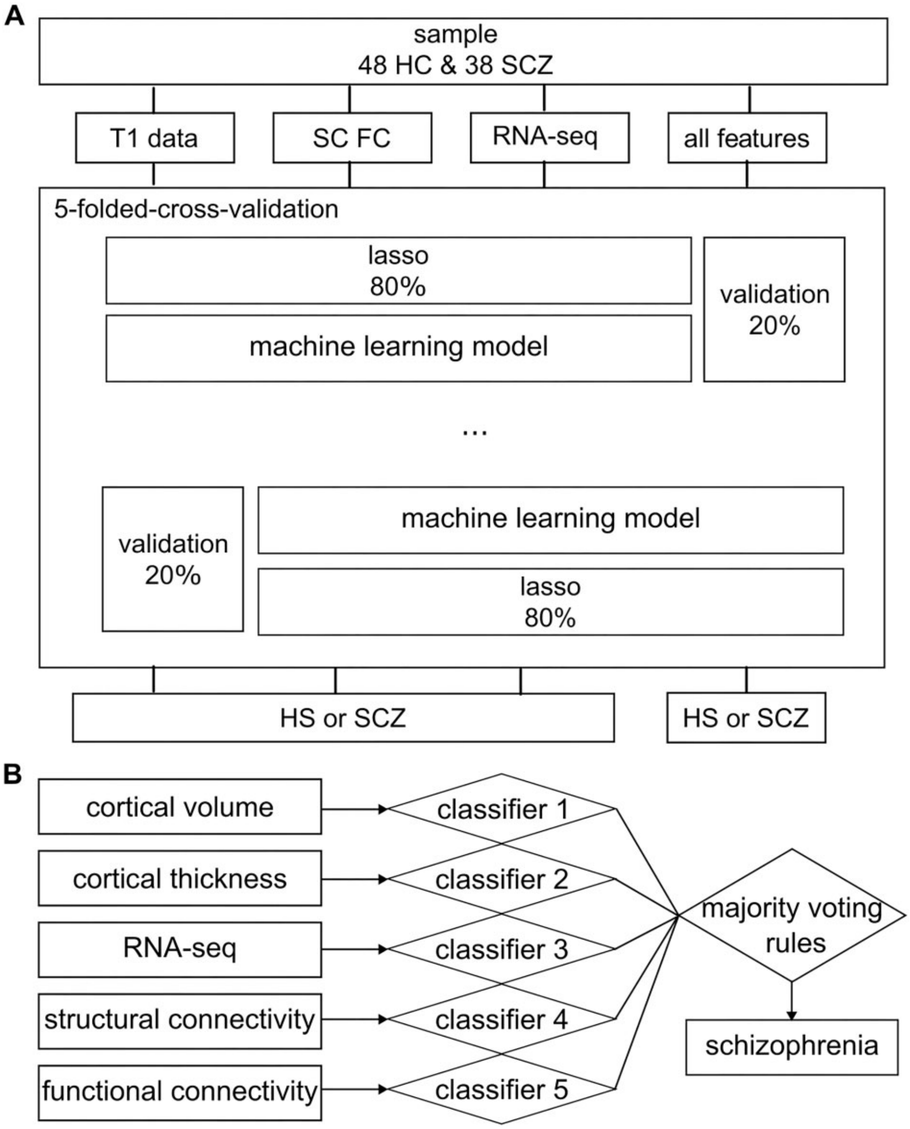

Figure 1: Scheme of feature fusion evaluation. (**A**) The first ML integration framework. The framework integrates five features into a multi-omics classification scheme and employs ML algorithms to classify SZ patients and HC controls after feature selection using the LASSO model in each fold cross-validation. (**B**) Majority voting rules. The second ML integration framework trains separate SVM classifiers on each modality and employs the LASSO model and 5-fold cross-validation for feature selection.

cortical thickness, RNA-seq, structural connectivity, and functional connectivity. These five feature matrices were scaled by subtracting the mean of each feature and were divided by the standard deviation, such that all features had a mean of zero and a standard deviation of one. Feature selection was performed using the least absolute shrinkage and selection operator (LASSO) model (Tibshirani, 1996) based on penalty terms in the embedding method. LASSO allows the selection of characteristics that are most important for the prediction of the target variable by constraining the regression coefficients given the predictor variables. The LASSO model is implemented by optimizing the loss function while minimizing the sum of the residual sum of squares and the penalty term. The loss function can be expressed as:

$$||y - X\beta||^2 + \lambda||\beta||_1$$

where y denotes the labels for SZ and HC, X is the macro brain image features or micro genetic data, β is the correlation coefficients, $||.||_1$ denotes the L1 paradigm, and λ is the regularization parameter.

Cross validation

Five-fold cross-validation was used, dividing the dataset into five subdivisions and taking four subdivisions each time as the training set and the remaining one sub-division as the test set. Stratified KFold (Widodo *et al.*, 2022) was used to perform 5-fold cross-validation in this study. To avoid feature leakage, feature selection was only applied to the training dataset. The trained classifier was tested on previously unseen test data samples.

Feature fusion in conventional ML

To examine whether integrating transcription and neuroimaging data boosts the performance in SZ classification, we evaluated two ML integration frameworks. The first ML framework directly concatenates the five features by splicing the features at input superposition. We implemented several ML approaches,

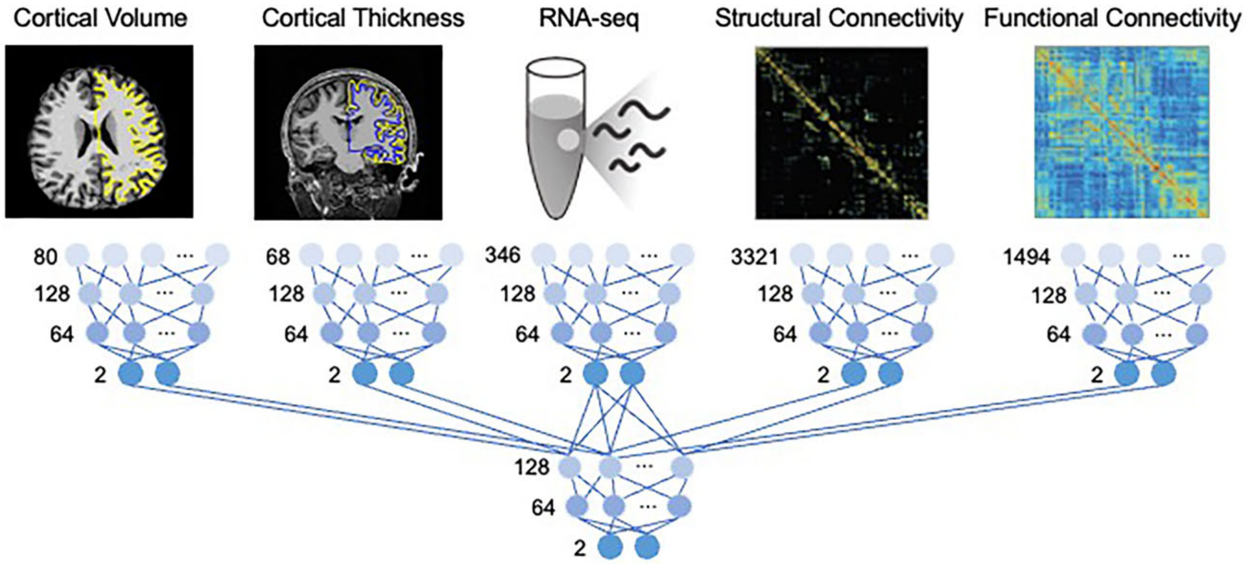

Figure 2: Neural network. The input nodes are based on the number of features, and the output nodes are based on the number of categories (HC, SZ). The sixth network used the output vectors of the first five networks as input.

including support vector machine (SVM), *k*-nearest neighbor algorithm (KNN), decision tree, and random forest, to identify SZ patients from HC (Fig. 1a). The parameters of the conventional ML model are described in the Supplementary Methods. The performance of these methods was assessed through the receiver operating characteristic curve, accuracy, precision, recall, specificity, and F1-score. We additionally assessed the effectiveness of the classification methodology by conducting 100 permutation tests on various multi-omics datasets, in which the labeled data were randomly permuted to show the robustness and dependability of the classification approach.

For the second ML framework, we trained a separate classifier on each modality and determined the final classification results based on the output of all classifiers. This framework also uses the LASSO model for feature selection and 5-fold cross-validation. Majority voting rules were used, which means a decision can only be made when more than half of classifiers agree (Fig. 1b).

Feature fusion in neural network

We also examined multi-omics data fusion using a neural network. A total of six DL networks were constructed based on cortical volume, cortical thickness, RNA-seq, SC, FC, and output of the first five DL networks. The output vector consists of the probability that the sample is classified as HC and SZ. The construction, training, and evaluation of the six deep-learning networks are shown in Fig. 2. The number of input nodes is based on the number of features, and the number of output nodes is based on the number of categories (i.e. HC, SZ). The connectivity matrix was reshaped into a one-dimensional vector with the upper triangles removed, and was stacked along the participants, resulting in a matrix with the dimensions of $n_{\text{subjects}} \times n_{\text{connections}}$ {i.e. 86 × [(82 × 81)/2]}. The input of the sixth network consists of the output vectors of the first five networks. Feeding each training data point through the network produces the output vector of weights. This output vector was compared to the target values, with any difference in the predicted outcome and the real outcome (i.e. HC, SZ) defined as an error using the cross-entropy error function. To simplify the model, the same number of knots was used for the hidden layers (second and third layers). The hidden layer nodes were chosen as 128 and 64, following the convention of using powers of 2. The result of setting the number of nodes in the hidden layer dependent on the input nodes is shown in Supplementary Table 5, which shows similar results compared to the original network. Errors were calculated after each training iteration. The training was stopped when the validation error ceased to decrease. After the training stage, the performance of the obtained neural network was assessed in the evaluation phase using the test dataset. The softmax activation function was used for the output nodes and the output node with the highest probability was selected as the predicted class label using a winner-take-all approach.

Results

Evaluation of multi-scale data fusion

We first illustrate the classification performance (i.e. AUC) of the integrated feature fusion in contrast to the performance achieved by each single modality (Fig. 3). The number of features in the classifier are 17, 19, 74, 150, 45, and 156, separately. Corresponding outcomes across these varying feature sets are detailed in Supplementary Figs. 2–4 and Supplementary Tables 4–6, showing similar performance to what we report in the main text. For all classifiers, the first multi-omics fusion model in conventional ML achieved an averaged AUC of 0.76–0.92, which outperformed any single-modality model (AUC 0.64–0.84) (Table 2). This result suggests that integrating multi-omics features enhances performance in SZ classification, with the AUC increased by 8.88–22.64%. Permutation testing (100 permutations) for the four ML methods showed the original AUC achieved in the multi-omics feature fusion model to significantly exceed random permutations ($P < 0.01$; mean AUC area of 0.50–0.52 and a variance of 0.01–0.02).

In addition to AUC, accuracy, precision, recall, specificity, and F1-score are presented in Table 3. For SVM and KNN, the accuracy, precision, recall, specificity, and F1-score of the first multi-omics fusion method in conventional ML showed better performance compared to any single modality. For the decision tree model, the accuracy achieved through multi-omics integration demonstrated superior performance. For the random forest model, the

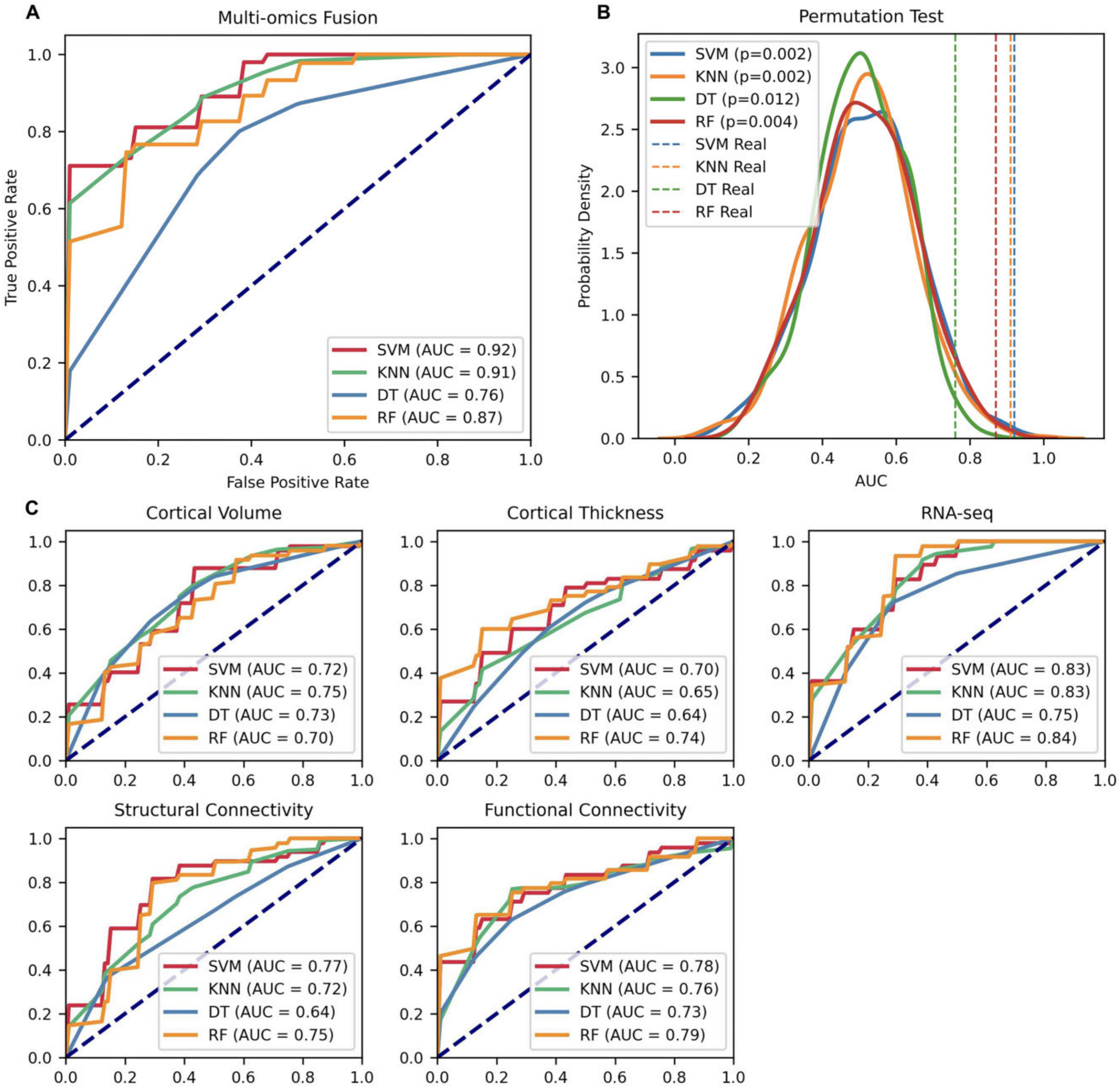

Figure 3: The first ML integration framework. (**A**) Multi-omics fusion in conventional ML. The first multi-omics model in conventional ML outperforms any single-omics model in terms of AUC, with an increase of 0.06–0.17. (**B**) Permutation test. Permutation testing (100 permutations) in four ML methods shows a significant difference between the original AUC and random permutations ($P < 0.05$). The mean AUC area is 0.50–0.52, and the variance is 0.01–0.02. (**C**) Single-omics model in conventional ML. DT, decision tree; RF, random forest.

Table 2: AUC of the first ML integration framework.

	SVM	KNN	DT	RF
Cortical volume	0.72	0.75	0.73	0.70
Cortical thickness	0.70	0.65	0.64	0.74
RNA-seq	0.83	0.83	0.75	0.84
Structural connectivity	0.77	0.72	0.64	0.75
Functional connectivity	0.78	0.76	0.73	0.79
Multi-omics fusion	0.92	0.91	0.76	0.87

DT, decision tree; RF, random forest.

accuracy was slightly lower than models solely based on RNA-seq data (Table 3).

Note that the performance of the first ML integration framework was better than the second ML integration framework. For the second ML integration framework, the accuracy, precision, recall, specificity, and F1-score were 65.10, 61.89, 97.78, 23.57, and 75.71%, respectively. Moreover, using the mean output from each classifier (predicted positive probability) for decision-making revealed similar results (Supplementary Table 7).

Contributions of different brain features

We further examined the contributions of different brain features in the first multi-omics integration framework using conventional ML. Correlation coefficients calculated by the LASSO model in 5-fold cross-validation were superimposed, which yielded 15 brain regions with high correlations (Fig. 4) (Scholtens *et al.*, 2021). These regions, including for example the right pallidum, left posterior cingulate, right frontal pole, and right temporal pole, were key regions that could identify SZ patients in our first multi-omics fusion model using conventional ML. The top 10 abnormal brain re-

Table 3: Accuracy, precision, recall, specificity, and F1-score of the first ML integration framework.

		Cortical volume (%)	Cortical thickness (%)	RNA-seq (%)	SC (%)	FC (%)	Multi-omics fusion (%)
SVM	Accuracy	72.09	67.45	79.15	77.97	73.14	80.20
	Precision	75.51	70.61	78.62	76.32	76.34	80.81
	Recall	79.11	73.33	87.56	89.56	76.22	84.89
	Specificity	63.93	60.36	68.57	63.21	67.50	73.21
	F1-score	75.59	71.21	82.36	82.17	74.65	82.23
KNN	Accuracy	61.63	66.21	82.55	69.74	73.20	75.62
	Precision	62.56	66.49	77.62	67.18	74.00	72.41
	Recall	81.33	81.56	98.00	93.78	82.67	91.78
	Specificity	37.86	47.50	62.86	40.00	60.00	55.71
	F1-score	70.18	72.87	86.46	77.70	76.72	80.65
DT	Accuracy	65.23	62.88	73.33	64.05	74.51	82.68
	Precision	69.01	68.77	79.81	69.22	77.40	85.59
	Recall	70.67	65.11	70.67	67.11	78.67	83.33
	Specificity	58.93	61.07	76.43	60.36	67.50	81.79
	F1-score	69.39	65.64	74.13	67.65	76.85	83.76
RF	Accuracy	66.34	60.59	79.02	76.80	73.20	75.56
	Precision	70.73	64.82	81.13	74.60	75.29	78.05
	Recall	74.89	65.11	83.33	89.56	78.44	78.44
	Specificity	56.43	55.71	73.57	61.07	65.00	70.71
	F1-score	71.50	64.08	81.35	81.27	75.70	77.21

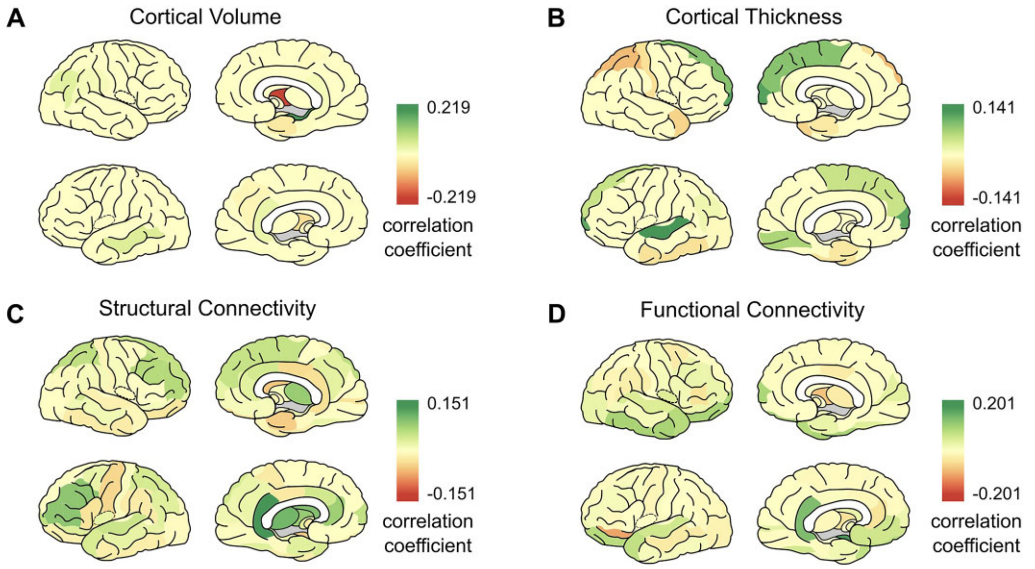

Figure 4: Contributions of features. (**A**) Abnormal brain regions of cortical volume. (**B**) Abnormal brain tegions of cortical thickness. (**C**) Abnormal brain regions of structural connectivity. (**D**) Abnormal brain regions of functional connectivity.

gions found in the previous four modalities and their correlation coefficients are shown in Supplementary Table 2.

Leave-one-feature-out framework.

We first illustrate the classification performance (i.e. AUC) of the leave-one-feature-out framework (Supplementary Fig. 1). For all classifiers, the leave one feature out framework model achieved an average AUC of 0.67–0.92 (Table 4), which was comparable to the first multi-omics integration framework using conventional ML (AUC 0.76–0.92) (Table 4).

Also note that when the RNA-seq data were removed, the performance of the model declined remarkably (e.g. AUC dropped to 0.75–0.82) (Table 4). This decline was consistent with the superior performance observed with RNA-seq data alone (AUC 0.75–0.84) (Table 2), underscoring the importance of transcriptome data.

Table 4: AUC of the leave-one-feature-out framework.

	SVM	KNN	DT	RF
Multi-omics fusion	0.92	0.91	0.76	0.87
Leave cortical volume	0.90	0.87	0.77	0.90
Leave cortical thickness	0.92	0.86	0.80	0.91
Leave RNA-seq	0.81	0.80	0.75	0.82
Leave structural connectivity	0.90	0.90	0.67	0.90
Leave functional connectivity	0.87	0.86	0.79	0.91

Table 5: Neural network results.

	AUC	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Cortical volume	0.58	57.14	63.64	38.89	76.47	48.28
Cortical thickness	0.58	57.14	53.57	88.24	27.78	66.67
RNA-seq	0.56	57.14	61.90	65.00	46.67	63.41
Structural connectivity	0.54	51.43	48.28	87.50	21.05	62.22
Functional connectivity	0.72	68.57	86.67	59.09	84.62	70.27
Multi-omics fusion	0.71	71.43	73.33	64.71	77.78	68.75

Evaluation of DL neural network

We also examined the performance of the DL neural network. The accuracy achieved for multi-omics feature fusion using neural network was 80.00%, which outperformed that of single modality DL networks (Table 5). This again confirms that the integration of imaging and transcriptomic data could improve the accuracy of SZ classification, regardless of the choice of the classification methodology.

Discussion

In the current study, we evaluated whether ML ensemble methods that combine transcriptomic and neuroimaging data could improve the accuracy of identifying patients with SZ. Specifically, the ML integration framework demonstrated a remarkable increase in average accuracy for multi-omics data. This increase ranges between 4.64 and 13.91% when compared to the single-omics average, accompanied by AUC improvements of 8.88 to 22.64%. Moreover, the DL network revealed a substantial enhancement in the average accuracy of multi-omics, which was 16.57% higher than that of the single-omics average. In summary, our findings provide empirical evidence for the choice of data fusion strategies when integrating multi-omics and trans-scale data to diagnose SZ.

Particularly in the first ML integration framework, our research provides methodological references for studies of SZ classification models, using algorithms such as SVM, decision tree, random forest, and KNN. The majority voting rule could not effectively improve the decision-making ability of the model. The reason might be that information is still learned from the single-omics features during classifier training, and the advantage of multi-omics feature learning is not taken into account. In the DL neural network, hidden layers could learn complex nonlinear relationships, demonstrating excellent recognition performance for HC and SZ. For all these methods, the addition of transcriptional data improved the predictability of SZ in contrast to simply using neuroimaging data, confirming that integrating transcriptomic data and neuroimaging enhances the effectiveness of ML/DL models.

The average AUC of the leave one feature out framework closely matches that of the multi-omics fusion model in conventional ML. Despite similar results between the two models, the study preserves all five features due to their unique contributory insights and the broader objective of pinpointing abnormal brain regions in patients with SZ beyond mere classification. Notably, the exclusion of RNA-seq data led to a remarkable decrease in model performance, reinforcing the critical role of integrating micro-omics data for a more comprehensive understanding of the disease.

Abnormal brain regions in patients with SZ were also found by the LASSO model. These findings were consistent with previous studies, demonstrating the abnormalities of the right pallidum, left isthmus cingulate, right frontal pole, and right temporal pole in individuals with SZ (van den Heuvel *et al.*, 2010; Schijven *et al.*, 2023). These results together guided feature selection for future studies that develop ML/DL diagnostic models for SZ.

Several considerations have to be drawn in the current study. First, the sample size is small, consisting of only 38 SZ and 48 HC participants. This limited sample size may hinder the comprehension of disease heterogeneity and the translation of models to new, unseen datasets. Future research on SZ would benefit from expanded samples. Second, the RNA-seq data derived from blood samples were used in the current study. Although gene expressions were correlated between blood and brain samples, our data still could not accurately reflect gene expression in specific brain areas that might be directly underlined alterations in the central nervous system. Third, the models we used for evaluation are relatively basic owing to the limitations inherent in the dataset. Future research could potentially explore more sophisticated and advanced models to improve the accuracy of disease progression prediction (Zhao *et al.*, 2020; Lei *et al.*, 2023; Yue *et al.*, 2023).

In conclusion, by combining macroscale brain imaging data and microscale gene transcriptome data, this study demonstrates the great potential and prospect of the application of multi-omics modeling in assisting and optimizing the diagnosis of SZ.

Supplementary data

Supplementary data is available at *PSYRAD* Journal online.

Author contributions

Mengya Wang (Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing), Shu-Wan Zhao (Writing – review & editing), Di Wu (Funding acquisition, Data curation), Ya-Hong Zhang (Data curation), Yan-Kun Han (Data curation), Kun Zhao (Writing – review & editing), Ting Qi (Writing – review & editing), Yong Liu (Writing – review & editing), Long-Biao Cui (Conceptualization, Funding acquisition, Data curation, Writing – review & editing), and Yongbin Wei (Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing)

Conflict of interest

None declared.

Acknowledgments

Y.W. received grants from the Beijing Municipal Natural Science Foundation (7232341) and the National Natural Science Foundation of China (82202264) during the conduct of the study. L.-B.C.

received a grant from the National Natural Science Foundation of China (82271949). D.W. received a grant from the Key Research and Development Program of Shaanxi Province (2023-YBSF-444). This study is supported by Super Computing Platform of Beijing University of Posts and Telecommunications.

References

Cao H, Zhou H, Cannon TD (2020) Functional connectome-wide associations of schizophrenia polygenic risk. *Mol Psychiatry* **26**:2553–61.

Chen S, Zhou Y, Chen Y, *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**:i884–90.

Chen Z, Liu X, Yang Q, *et al.* (2023) Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: a systematic review. *JAMA Netw Open* **6**: e231671.

Cui L-B, Wei Y, Xi Y-B, *et al.* (2019) Connectome-based patterns of first-episode medication-naïve patients with schizophrenia. *Schizophr Bull* **45**:1291–9.

Cui L-B, Zhao Shu-Wan, Zhang Ya-Hong, *et al.* (2023)Multi-omic transcriptional, brain, and clinical variations in schizophrenia. medRxiv.10.1101/2023.05.30.23290738

De Lange SC, Helwegen K, Van Den Heuvel MP (2021) Structural and functional connectivity reconstruction with CATO—a connectivity analysis toolbox. *Neuroimage* **273**:120108.

De Reus MA, Van Den Heuvel MP (2013) Estimating false positives and negatives in brain networks. *Neuroimage* **70**:402–9.

Desikan RS, Ségonne F, Fischl B, *et al.* (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**:968–80.

Fischl B (2012) FreeSurfer. *Neuroimage* **62**:774–81.

Gao S, Calhoun VD, Sui J (2018) Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci Ther* **24**:1037–52.

Gao Z, Xiao Y, Zhu F, *et al.* (2023) The whole-brain connectome landscape in patients with schizophrenia: A systematic review and meta-analysis of graph theoretical characteristics. *Neurosci Biobehav Rev* **148**:105144.

Griffa A, Baumann PS, Klauser P, *et al.* (2019) Brain connectivity alterations in early psychosis: from clinical to ne uroimaging staging. *Transl Psychiatry* **9**:62.

Guggenmos M, Schmack K, Veer IM, *et al.* (2020) A multimodal neuroimaging classifier for alcohol dependence. *Sci Rep* **10**:298.

Hulshoff Pol HE, Schnack HG, Bertens MGBC, *et al.* (2002) Volume changes in gray matter in patients with schizophrenia. *Am J Psychiatry* **159**:244–50.

Jenkinson M, Beckmann CF, Behrens TEJ, *et al.* (2012) FSL. *Neuroimage* **62**:782–90.

Ji Y, Zhang X, Wang Z, *et al.* (2021) Genes associated with gray matter volume alterations in schizophrenia. *Neuroimage* **225**:117526.

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**:357–60.

Lam M, Chen C-Y, Li Z, *et al.* (2019) Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet* **51**:1670–8.

Lei B, Zhu Y, Yu S, *et al.* (2023) Multi-scale enhanced graph convolutional network for mild cognitive impairment detection. *Pattern Recognit* **134**:109106.

Scholtens L, de Lange S, van den Heuvel M (2021) Simple brain plot. *Zenodo*. 10.5281/zenodo.5346593

Liu N, Xiao Y, Zhang W, *et al.* (2020) Characteristics of gray matter alterations in never-treated and treated chronic schizophrenia patients. *Transl Psychiatry* **10**:136.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**:550.

Mccutcheon RA, Reis Marques T, Howes OD (2020) Schizophrenia-an overview. *JAMA Psychiatry* **77**:201–10.

Morez J, Sijbers J, Vanhevel F, *et al.* (2021) Constrained spherical deconvolution of nonspherically sampled diffusion MRI data. *Hum Brain Mapp* **42**:521–38.

Mori S, Crain BJ, Chacko VP, *et al.* (1999) Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Ann Neurol* **45**:265–9.

Romme IAC, De Reus MA, Ophoff RA, *et al.* (2017) Connectome disconnectivity and cortical gene expression in patients with schizophrenia. *Biol Psychiatry* **81**:495–502.

Rothberg JM, Hinz W, Rearick TM, *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**:348–52.

Sadeghi D, Shoeibi A, Ghassemi N, *et al.* (2022) An overview of artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: methods, ch allenges, and future works. *Comput Biol Med* **146**:105554.

Schijven D, Postema MC, Fukunaga M, *et al.* (2023) Large-scale analysis of structural brain asymmetries in schizophrenia via the ENIGMA consortium. *Proc Natl Acad Sci USA* **120**: e2213880120.

Sui J, Zhi D, Calhoun VD, *et al.* (2023) Data-driven multimodal fusion: approaches and applications in psychiatric research. *Psychoradiology* **3**:1–19.

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *J R Stat Soc Series B Stat Methodol* **58**:267–88.

Trubetskoy V, Pardiñas AF, Qi T, *et al.* (2022) Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**:502–8.

Van Den Heuvel MP, Mandl RCW, Stam CJ, *et al.* (2010) Aberrant frontal and temporal complex network structure in schizophrenia: a graph theoretical analysis. *J Neurosci* **30**:15915–26.

Van Der Meer D, Shadrin AA, O'Connell K, *et al.* (2022) Boosting schizophrenia genetics by utilizing genetic overlap with brain morphology. *Biol Psychiatry* **92**:291–8.

Watanabe K, Taskesen E, Van Bochoven A, *et al.* (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**:1826.

Wei Y, De Lange SC, Savage JE, *et al.* (2023) Associated genetics and connectomic circuitry in schizophrenia and bipolar disorder. *Biol Psychiatry* **94**:174–83.

Widodo S, Brawijaya H, Samudi S (2022) Stratified K-fold cross validation optimization on machine learning for prediction. *Sinkron* **7**:2407–14.

Yue G, Wei Peishan, Zhou Tianwei, *et al.* (2023) Specificity-aware federated learning with dynamic feature fusion network for imbalanced medical image classification. *IEEE J Biomed Health Inf*, 10.1109/JBHI.2023.3319516

Zhao C, Wang T, Lei B (2020) Medical image fusion method based on dense block and deep convolutional generative adversarial network. *Neural Comput Appl* **33**:6595.

Received: 4 December 2023; **Revised:** 26 February 2024; **Accepted:** 25 March 2024