Received: 4 August 2023 | Revised: 26 April 2024 | Accepted: 12 July 2024

DOI: 10.1002/hbm.26799

TECHNICAL REPORT

WILEY

Imaging-genomic spatial-modality attentive fusion for studying neuropsychiatric disorders

Md Abdur Rahaman[1,2] | Yash Garg[3] | Armin Iraji[2,4] | Zening Fu[2] | Peter Kochunov[5] | L. Elliot Hong[5] | Theo G. M. Van Erp[6,7] | Adrian Preda[8] | Jiayu Chen[2] | Vince Calhoun[1,2]

[1]Georgia Institute of Technology, Atlanta, Georgia, USA

[2]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Atlanta, Georgia, USA

[3]Nokia Bell Labs, Murray Hill, New Jersey, USA

[4]Department of Computer Science, Georgia State University, Atlanta, Georgia, USA

[5]University of Maryland Center for Brain Imaging Research, College Park, Maryland, USA

[6]Clinical Translational Neuroscience Laboratory, Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, California, USA

[7]Center for the Neurobiology of Learning and Memory, University of California Irvine, Irvine, California, USA

[8]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, California, USA

Correspondence
Md Abdur Rahaman, University of Maryland Center for Brain Imaging Research, College Park, MD, USA.
Email: mrahaman8@gatech.edu

Funding information
National Science Foundation, Grant/Award Number: 2112455; National Institutes of Health, Grant/Award Number: R01MH118695

Abstract

Multimodal learning has emerged as a powerful technique that leverages diverse data sources to enhance learning and decision-making processes. Adapting this approach to analyzing data collected from different biological domains is intuitive, especially for studying neuropsychiatric disorders. A complex neuropsychiatric disorder like schizophrenia (SZ) can affect multiple aspects of the brain and biologies. These biological sources each present distinct yet correlated expressions of subjects' underlying physiological processes. Joint learning from these data sources can improve our understanding of the disorder. However, combining these biological sources is challenging for several reasons: (i) observations are domain specific, leading to data being represented in dissimilar subspaces, and (ii) fused data are often noisy and high-dimensional, making it challenging to identify relevant information. To address these challenges, we propose a multimodal artificial intelligence model with a novel fusion module inspired by a bottleneck attention module. We use deep neural networks to learn latent space representations of the input streams. Next, we introduce a two-dimensional (spatio-modality) attention module to regulate the intermediate fusion for SZ classification. We implement spatial attention via a dilated convolutional neural network that creates large receptive fields for extracting significant contextual patterns. The resulting joint learning framework maximizes complementarity allowing us to explore the correspondence among the modalities. We test our model on a multimodal imaging-genetic dataset and achieve an SZ prediction accuracy of 94.10% (p < .0001), outperforming state-of-the-art unimodal and multimodal models for the task. Moreover, the model provides inherent interpretability that helps identify concepts significant for the neural network's decision and explains the underlying physiopathology of the disorder. Results also show that functional connectivity among subcortical, sensorimotor, and cognitive control domains plays an important role in characterizing SZ. Analysis of the spatio-modality attention scores suggests that structural components like the supplementary motor area, caudate, and insula play a significant role in SZ. Biclustering the attention scores discover a multimodal cluster that includes genes CSMD1, ATK3, MOB4, and HSPE1, all of which have been identified as relevant to SZ. In summary, feature attribution appears to be especially useful for probing the transient and confined but decisive patterns of complex disorders, and it shows promise for extensive applicability in future studies.

KEYWORDS
bottleneck attention module (BAM), functional network connectivity (FNC), imaging-genetics, multimodal fusion, resting-state functional and structural MRI, schizophrenia classification, single-nucleotide polymorphism (SNP)

1 | INTRODUCTION

Our brain processes multimodal signals from the outer world to understand real events and respond accordingly (Bayoudh et al., 2021; Guo et al., 2019; Kosslyn et al., 2010; Nanay, 2018; Summaira et al., 2021). Leveraging information from diverse sources allows for a better understanding of a given phenomenon. Likewise, multimodal learning enables researchers and practitioners to benefit from the unique strengths of each modality, leading to improved performance, enhanced accuracy, and richer insights (Ngiam et al., 2011; Sohn et al., 2014). Applications of multimodal learning models are wide-ranging and include fields such as computer vision (Hosseinzadeh & Wang, 2021; Liu et al., 2021; Xi et al., 2020), healthcare (Huang, Pareek, Seyyedi, et al., 2020; Menon & Krishnamurthy, 2021; Naderi et al., 2019), and speech recognition (Palaskar et al., 2018; Yuhas et al., 1989), among others (Sengupta et al., 2020). Moreover, the potential applications of multimodal learning in studying mental disorders are vast and continue to expand (Calhoun & Sui, 2016; Rahaman et al., 2021; Rahaman et al., 2023; Shi et al., 2017; Venugopalan et al., 2021; Zhang et al., 2022). The rationale lies in multiple biological domains that are affected by the underlying physical conditions. Therefore, maximizing the complementarity among these sources can potentially lead to a more comprehensive understanding of the data.

Research indicates that useful biological information is encoded across different sources. Neuroimaging provides structural and functional information about the brain through various imaging modalities (Calhoun & Sui, 2016; Zhang et al., 2022). Genetic data provide another source of information regarding disease-related aberrations (Hardoon et al., 2009; Le Floch et al., 2012). The integration of data from different modalities and sources can offer a richer and more nuanced understanding of complex mental disorders. Data from multiple modalities are not mutually exclusive but complement each other in describing brain processes (Zhang et al., 2022). Particularly, neuroimaging modalities, when combined, can achieve enhanced temporal and spatial resolution and bridge the gap between physiological and cognitive representations (Liu, Sun, & Zhang, 2018). Hence, multimodal learning frameworks have emerged as effective tools for analyzing data from multiple sources, including neuroimaging (Aine et al., 2017; Calhoun & Sui, 2016; Shi et al., 2017; Tulay et al., 2019) and genetic data (Bogdan et al., 2017; Rahaman et al., 2021). Past research has demonstrated significant correlations between structural and functional changes in the brain and mental disorders (Salgado-Pineda et al., 2011; Segall et al., 2012). Moreover, existing scientific literature points to a promising area of exploration: the correlation between genetic variants and neural activity concerning neuropsychiatric disease-related degeneration (Hardoon et al., 2009; Le Floch et al., 2012; Liu & Calhoun, 2014). Such a disorder schizophrenia (SZ) is genetically complex and affects the brain's structure and function (MacDonald III et al., 2009; Meyer-Lindenberg, 2010; Rahaman et al., 2021). Nevertheless, significant challenges arise in the joint analysis of genetic, structural, and functional data, as they often carry information at different scales and formats. A perceptive fusion module is necessary for enhancing the model's performance as it ensures the judicious use of the most informative sources in subsequent tasks. A bottleneck strategy might facilitate the integration of these diverse knowledge domains, overcoming the obstacles posed by differing data scales and formats. The bottleneck in a neural network (NN) is a layer with fewer neurons than the layer below or above it (He et al., 2016; Park et al., 2018). It helps to learn representations better and emphasize salient features for the target variable. Thus, our intuition is to operate fusion in this layer with bottleneck attention.

Research has provided a thorough classification of fusion methodologies (Huang, Pareek, Zamanian, et al., 2020; Poria et al., 2017). These fusion strategies are broadly divided into three subcategories: early fusion, mid/intermediate fusion, and late fusion, depending on the phase of the model where the integration takes place. The mid-fusion approach is particularly notable due to its wide range of applications. In this scheme, fusion occurs after feature extraction from different input modes (Huang, Pareek, Zamanian, et al., 2020). However, integrating multiple sources in the intermediate phase presents a challenge, as differences in data types, collection methods, scales, and preprocessing can lead to inconsistencies in data representations. To address this issue, modality-specific NNs are used to first learn the hidden space representations of the inputs. These networks translate the multimodal inputs into a uniform embedding space, after which the mid-fusion module is used for the integration (Poria et al., 2017; Roitberg et al., 2019). It is important to note that even when employing latent space fusion, the resulting fused feature map can be high-dimensional, with each unit's contribution to the subsequent task potentially varying in importance. Consequently, simply connecting the fused tensor to the predictor may limit the model's performance. This problem is especially pronounced when data modalities are unevenly informative about the downstream task, a common occurrence in medical data collection. Neuroimaging and genomics datasets often contain weak descriptors of the underlying biology for a few samples. Therefore, the fusion module needs to enhance the synergies between these data sources. Previous research has proposed various embedding fusion techniques, including attention-based (Hazarika et al., 2018; Vaswani et al., 2021), multiplicative combination layer (Liu, Li, et al., 2018), and transformer (Nagrani et al., 2021). An early adaptation of the bottleneck attention module (BAM) (Park et al., 2018) known as the mBAM (Rahaman et al., 2022) examined both spatial and modality dimensions. It employed a fully connected (FC) layer with compression to learn spatial attention from the fused two-dimensional tensor.

In this study, we present a fusion module called spatio-modality fusion using bottleneck attention, which carefully examines the amalgamation. We use dilated convolutional (Yu & Koltun, 2015) methods to learn the contextual pattern. The module explores spatial dimensions using a convolutional neural network (CNN) (Gu et al., 2018) augmented by a larger receptive field (RF), a feature known as dilated convolution. Using dilated convolutions significantly expands the RF, facilitating the collection of contextual patterns from the combined data (Park et al., 2018). These contextual patterns play a crucial role in the downstream tasks. Our study shows that the dilated convolutional layer performs better than the FC layers (Figure 3). In earlier studies, the FC layers lost considerable spatial information while downsampling from the combined tensor to a one-dimensional vector (Rahaman et al., 2022). Furthermore, the subsequent attention operation performed on the compressed version produced a low-dimensional (vector) mask. Our method, on the other hand, implements spatial attention on the original version of the fused tensor, generating a two-dimensional attention mask. Simultaneously, the module learns to apply attention across the modalities to select the best source, and in spatial dimensions to mask the compressed feature vectors, fostering a richer knowledge extraction. The attention mask delineates the significance of each feature on classification, in our case identifying SZ. Therefore, the mask can also be utilized to generate other analytics of the test samples, for example, subgrouping based on disease relevance.

The modalities we use for the classification are structural and functional neuroimaging data and genome-wide polymorphism data. We test the model on a dataset comprising 437 subjects, including individuals with SZ (162) and controls (275), intending to classify SZ. Our proposed method produces a multidimensional attention mask to elucidate the model's decisions and underlying neurobiological basis. This attention mask encodes the relative importance of each modality and spatial significance. The spatio-modality attention identifies structural magnetic resonance imaging (sMRI) components—such as the supplementary motor area (SMA), left insula, caudate, and temporal pole—of high importance for detecting SZ (Figure 4). The attention scores on static functional connectivity suggest that several connections among the sensorimotor (SM), subcortical (SC), and cognitive control (CC) are particularly salient in SZ. Biclustering the attention scores from three modalities discovers a multimodal cluster that includes a subset of relevant structural components, functional connections, and genes. The implicated genes, CSMD1, ATK3, MOB4, and HSPE1, have been previously recognized as relevant to SZ. The primary features of our method are as follows:

1. A fusion module capable of encoding both modality and spatial attention.
2. A dilated convolution to extract spatial patterns with a large RF.
3. A two-dimensional spatio-modality attention mask, enabling further data analytics such as subgroup identification.
4. A self-explainable model via the attention scores for modality and contextual dimension.

2 | DATA PREPROCESSING

2.1 | Structural MRI

We preprocessed sMRI scans using statistical parametric mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm/). The preprocessing steps include unified segmentation and normalization of sMRI scans into gray matter, white matter, and CSF. During the segmentation step, we use a modulated normalization algorithm to generate gray matter volume. Following this, we use a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm to smooth the gray matter densities.

2.2 | Functional MRI

We use the SPM12 toolbox for preprocessing functional MRI data. To ensure steady-state magnetization, we remove the first five time points of the fMRI. Next, we carry out rigid body motion correction using the INRI-Align robust M-estimation approach and apply the slice-timing correction. We then spatially normalized fMRI images into the Montreal Neurological Institute standard space, using an echo-planar imaging template, and resampled to 3 × 3 × 3 mm^3 isotropic voxels. The images are then smoothed with a Gaussian kernel with FWHM = 6 mm, similar to the sMRI data.

2.3 | Genomics

The preprocessing steps for the genetic data are described in our prior work (Adhikari et al., 2019; Chen et al., 2013). There are several standard preprocessing tools for genomics data, and we use Plink (Purcell et al., 2007) for both preprocessing and imputation. Linkage disequilibrium (LD) pruning was administered at $r^2 < .9$. We use the Psychiatric Genomics Consortium (PGC) for SZ suggested genome-wide association study (GWAS) (Purcell et al., 2007) score to select the features. The analysis selects 1280 single-nucleotide polymorphisms (SNPs) distributed across 108 risk loci. The PGC study (He et al., 2015) reveals these SNPs express statistically significant associations with SZ at $p < 1 \times 10^{-4}$. The datasets verify the test retest experiment for MRI acquisitions by recording multiple repetitive scans for each subject and also account for the stable MRI signal.

More details about preprocessing and quality control are described in these studies (Fu, Iraji, et al., 2021; Iraji et al., 2022).

3 | OUR PROPOSED MULTIMODAL ARCHITECTURE

Figure 1 demonstrates our proposed architecture for multimodal learning. The model incorporates three modality-specific NNs, referred to as subnetworks, which are used for learning the latent space embedding from each input source. These NNs are selected based on their effectiveness given the type of input data, and we empirically validate their efficacy. The subnetworks conduct dimensionality reduction and account for missing entries and other discrepancies to effectively learn the representations from each incoming modality. The latent embeddings from all modalities are concatenated into a multidimensional tensor. This tensor goes through a novel spatio-modality attention-based fusion module, which is described in Figure 2. Once this is completed, the two-dimensionally attended fused embedding is used as input to a multilayer perceptron (MLP). This is followed by a SoftMax layer for classification. This approach allows us to extract rich features from multiple modalities and harness the power of NNs to classify complex data effectively.

3.1 | Input features quantification

This phase of the framework includes a few processing units for refining and performing the initial decomposition of the data collected from multiple biological domains. We execute a fully automated spatially constrained group ICA (GICA) using the Group ICA of fMRI

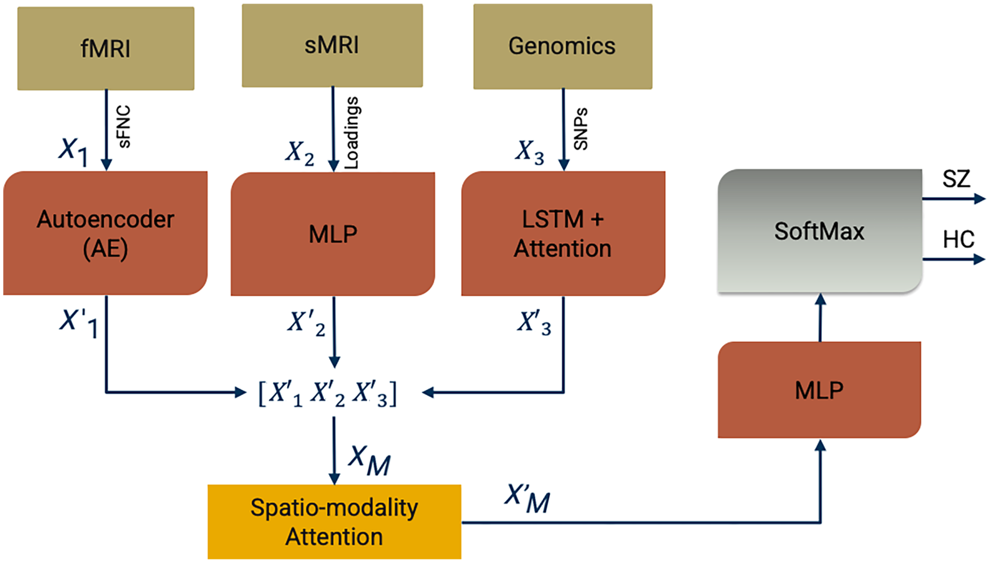

FIGURE 1 Our proposed multimodal framework with spatio-modality temporal attention. These models incorporate three major processing steps: Imaging-genomics preprocessing and dimensionality reduction, neural subnetworks for learning latent space embedding, and the predictor. We run group ICA on sMRI and fMRI data. We generate static functional network connectivity (sFNC) among the ICNs extracted from fMRI decomposition. We select the ICA loading parameter as the input feature for sMRI modality. The GWAS-based genetic variable selection is carried out for genetic modality. The subnetworks are deep neural networks for learning the modality-wise representation. Subnetwork 1 is an autoencoder (AE) for learning sFNC, subnetwork 2 is a multilayered perceptron (MLP) for sMRI loadings, and subnetwork 3 is a bi-directional long short-term memory (LSTM) unit with attention for learning SNPs. The combined embedding is attended in spatial and modality direction and sent through an MLP followed by a SoftMax layer for the classification. The architecture is jointly trained using an Adam optimizer.

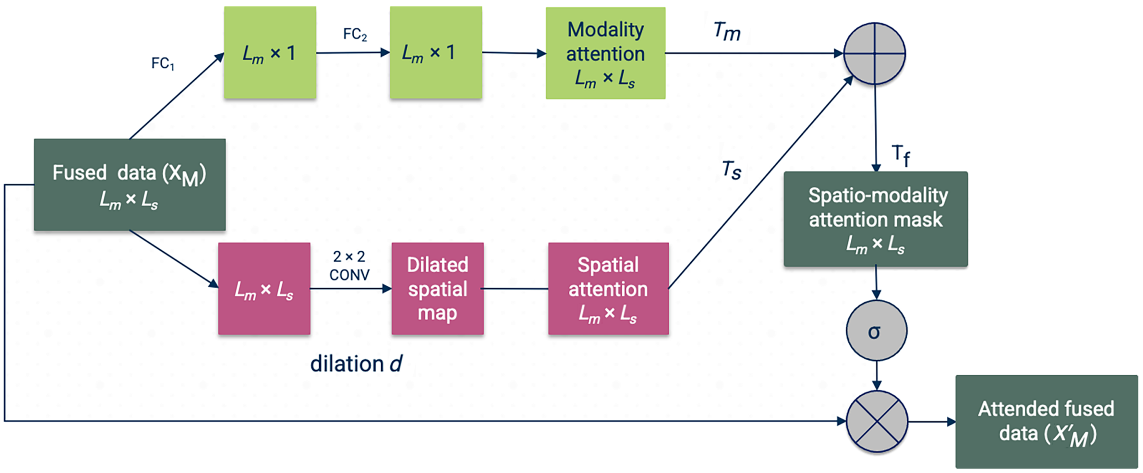
FC₁
$L_m \times 1$
FC₂
$L_m \times 1$
Modality attention $L_m \times L_s$
T_m
Fused data (X_M) $L_m \times L_s$
T_s
T_f
Spatio-modality attention mask $L_m \times L_s$
$L_m \times L_s$
2 × 2 CONV
Dilated spatial map
Spatial attention $L_m \times L_s$
dilation d
σ
Attended fused data (X'_M)

FIGURE 2 Our proposed spatio-modality attention module for multimodal fusion. The concatenated embeddings are sent through two different branches. (i) The modality branch that learns the cross-modality interactions and mounts it into T_m attention mask and (ii) the spatial branch (T_s) captures the relevant context from each biological site. These two masks are merged into a final attention mask T_f. The spatial branch uses dilated convolution for learning the contextual understanding of the multimodal tensor and fully connected layer for modality attention.

Toolbox (GIFT) available at this link: (http://trendscenter.org/software/gift) (Iraji et al., 2021) and the NeuroMark (Adhikari et al., 2019) template on the combined subjects from three datasets, which include FBIRN (Keator et al., 2016), COBRE (Aine et al., 2017), and MPRC (Adhikari et al., 2019). The selection of intrinsic component networks (ICNs) in this study is based on the NeuroMark template. We selected 53 pairs of ICNs and arranged them into seven functional networks based on prior anatomical and functional knowledge fields (Fu, Iraji, et al., 2021; Iraji et al., 2022). The total number of connectivity networks extracted was 53, covering the whole brain. These ICNs were distributed into seven functional domains (Fu, Iraji, et al., 2021; Fu, Sui, et al., 2021): sub-cortical (SC), auditory, SM, visual (VS), cognitive-control (CC), default-mode, and cerebellar domain. After estimating subject-specific networks, we compute the subject-wise static functional network connectivity (sFNC). The square matrix (53 × 53) represents the Pearson correlation between the time course of ICNs. We vectorize the sFNC matrix using the upper diagonal entries to ease the encoder's training process. A similar approach was used for structural MRI source-based morphometry (ICA on gray matter maps) (Gupta et al., 2019; Saha et al., 2022), resulting in 30 structural components along with their loading values. The provided ICA priors are estimated from the 6500 subjects used in the referred study (Abrol et al., 2017). The ICA on sMRI yields individual-level structural networks and corresponding loadings. For the sMRI modality, we used the loadings of the structural networks as the structural features. The third modality includes the set of SNPs selected from genomic data based on GWAS-significant SZ risk SNPs identified by the large PGC study.

3.2 | Deep neural networks

We employ distinct deep neural networks (DNNs) for extracting modality-wise features. The design choices are explained in the following subsections, and their efficacy is tested on the dataset. We use an autoencoder (AE) (Goodfellow et al., 2016) to learn the representation of sFNC. It follows an encoder and decoder architecture, with each submodule consisting of five linear layers. The encoder compresses the input and generates a compressed embedding. The decoder reconstructs the input from the encoded features map. The loss function computes the reconstruction loss as the mean square error (MSE). We use the sFNC matrix as an input from the fMRI modality and employ an AE with rectified linear unit activation for learning the representations. AE is effective for learning semantic meanings and compressed abstraction of input data with successful applications in diverse fields of study (Chen et al., 2017; Hong et al., 2015). We use Xavier initialization (Kumar, 2017) from Pytorch to set the initial values of the layers in the network. We applied a dropout strategy with a probability p of 0.2 to minimize overfitting. Overfitting refers to a situation where a model performs well on the training data but fails to generalize to unseen data. Dropout is a regularization technique that helps to mitigate this issue by randomly ignoring a subset of features during training, reducing the complexity of the model and promoting generalization. For the sMRI subnetwork, we use the loading parameters from GICA. The feature vector has a length of 30, where each value represents the expression level of a subject on a group structural independent component. In general, loadings are just betas/coefficients of the variance mixture linear equations (Calhoun & Allen, 2013). The loading value corresponds to how much variance a subject contributes to a given group component. As the loadings are the vector of discrete values, for simplicity, we deployed an MLP to efficiently learn the sMRI features. An MLP is a type of artificial NN that consists of multiple layers of nodes in a directed graph, with each layer FC to the next one. It can model complex, nonlinear relationships between input and output data, making it a suitable choice for feature extraction in this context. Each layer of the MLP takes the output of the previous layer (or the input data for the first layer), applies a set of weights (learned during training), and then applies an activation function, producing the output for that layer. By adjusting the weights during training, the MLP can learn to extract salient features from the input data that are informative for the task at hand, which in this case, is the classification of SZ. The subnetwork is FC since the data dimension (30) is lower than the other modalities. In the final layer of our network, we dilated the embeddings to 100 to ensure consistency with the size of latent features derived from different modalities. We implemented a bi-directional long short-term memory (bi-LSTM) network with an attention mechanism (Ashish, 2017) to extract features from SNPs. The use of a bi-LSTM was informed by its anticipated ability to capture contextual information from a sequence (Li et al., 2019; Melamud et al., 2016). We assume that neighboring SNPs may show LD with one other, potentially forming a neighborhood substructure. The bi-LSTM is designed to capture the localized semantics of genomic data, which might help differentiate between cases of the disorder and controls. In addition, the attention scores signify contributing neighbors to the context, further enhancing the DNN's discriminative power.

3.3 | Spatio-modality attention

We propose a fusion module that takes a multimodal joint embedding and probes the concatenated tensor in both modality and spatial dimensions. The architecture is demonstrated in Figure 2. The architecture is inspired by the BAM (Park et al., 2018), which has been implemented for learning channels and spatial attention in classification and prediction tasks (Tang et al., 2021; Woo et al., 2018; Yaseen et al., 2022). We adapt the module for fusing different modalities. The module is integrated with a mid/intermediate fusion (Boulahia et al., 2021) in a multimodal learning framework. In mid-fusion, the modality-specific DNNs generate compressed representations of input streams (bottleneck layers) from each source. Our intuition is to run two-dimensional attention on these bottleneck layers. Our module uses two simultaneous attention branches that forage significant bits from the fused tensor. These branches can be treated as two distinct scoring functions across modality and spatial dimensions. We use FC linear layers in modality dimensions to learn the attention weights per modality. FC layer, is a type of layer used in NNs where all neurons in the previous layer are connected to the neurons in the next layer. The FC operation is defined in the following Equation (2).

$$\mathrm{FC}(x) = W^T x + b \tag{1}$$

where $W \in \mathbb{R}^{N \times 1}$ and $b \in \mathbb{R}^{N \times 1}$ are the weight matrix and bias, respectively, and N is the input dimension. x is the input to the FC operation, in the context of Equation (1), and b is the bias vector in the FC layer.

For another branch, the module employs dilated convolution layers for extracting the relevant spatial patterns. The dilated convolutional filters arbitrate a large RF for collecting contextual information. In our study, we are using only three modalities, so we skip the reduction. However, the reduction ratio in modality direction could potentially help in combining a large number of modalities in further studies.

3.3.1 | Modality attention scoring (T_m)

The modality attention branch assigns scores to the modality dimension to signify the most informative and discriminative source for the downstream task. The modality attentions mask (T_m) gathers importance scores that characterize the contribution of input modality to the classification of SZ. Figure 2 illustrates the attention-weighting architecture. An FC linear layer (FC_1) is applied to the input tensor, X_M ($L_m \times L_s$) to reduce the dimension to the number of modalities (L_m). Here, L_s is defined as the latent embedding size consistent across the modalities. The compressed data are then passed through another FC layer (FC_2) to compute the attention weights for each modality. We also use batch normalization (BN) (Ioffe & Szegedy, 2015) to adjust the scale with the spatial branch. Moreover, BN improves the speed, performance, and stability of the NNs. Then, the tensor is expanded to the shape of the input tensor $L_\mathrm{m} \times L_\mathrm{s}$. We can formulate the operation as follows:

$$T_m = \mathrm{BN}\left(\mathrm{FC}_2\left(\mathrm{FC}_1(X_M)\right)\right) \tag{2}$$

Here, T_m refers to the transformed tensor after the application of BN and two FC layers.

3.3.2 | Spatial attention branch (T_s)

We opt for capturing the context from the aggregated multimodal embedding tensor, which extracts contextual information in the form of spatial patterns. For spatial attention, the submodule uses a CNN. We use dilated convolution (Yu & Koltun, 2015) to create a large RF that assists in apprehending the context. The dilated convolution inflates the kernel by inserting holes between the kernel elements (Yu & Koltun, 2015). An additional parameter, dilation ratio (d), indicates the extent to which the kernel is widened. There are usually d-1 spaces inserted between kernel entries. We use an empirically validated convolutional filter of dimension 2 × 2 with a dilation ratio (d) = 2 on the combined tensor. Figure 2 shows the dilated convolution for spatial significance. The branch weighs the fused data for identifying salient loci relevant to downstream tasks. The spatial attention mask is also expanded to a shape of $L_\mathrm{m} \times L_\mathrm{s}$. Equation 3 is the operation for spatial attention and illustrates the overall processing in the spatial branch.

$$T_s = \mathrm{BN}\left(d^{2\times 2}\left(d^{2\times 2}(X_M)\right)\right) \tag{3}$$

Here, T_s represents the spatial attention mask. The operation helps to create a large RF that captures the context or spatial pattern in the input data. We merge the spatial attention mask, T_s and the modality attention mask, T_m to generate the final mask, T_f (see Equations 4).

$$T_f = \sigma(T_s + T_m) \quad (4)$$

The operation in Equation (4) is an element-wise summation between T_s and T_m to compute the final fused attention mask, T_f. We apply a sigmoid function (σ) to bind the values of T_f between 0 and 1. The sigmoid function is commonly used in NNs to introduce nonlinearity and to map any input to a value between 0 and 1, making it especially suitable for models where we have to predict probabilities.

Next, we use element-wise multiplication operation ($\otimes$) to fuse the attention mask T_f and the input tensor X (Equation 5).

$$X_{\triangle} = X \otimes T_f \quad (5)$$

This operation highlights the important features in the input tensor according to the attention mask. The result is a new tensor matrix, which emphasizes the regions of the input that are most informative for the task at hand. We use $X_{\triangle}$ as the input to the MLP followed by a SoftMax layer to classify the samples (Figure 1). Together, these operations help to guide the model to focus on the most informative features across different modalities and spatial locations when performing the downstream task, such as classification or regression.

3.4 | Multimodal joint training

We split the dataset into two primary categories: 80% for training and 20% for evaluation. The evaluation data are then further divided equally into testing and validation sets. We train the joint model for 450 training epochs and validate its performance on the validation set. For the training phase, we use the *Adam* optimizer from Pytorch with a learning rate of 10^{-4} and a batch size of 32. The loss terms include MSE for the AE subnetwork reconstruction loss and binary cross-entropy for the model's classification loss. We implement early stopping to balance the training and validation loss, which eventually regularizes the model. Our training scheme optimizes for accuracy and saves the best-performing model. For the validation phase, we used the k-fold cross-validation (Bengio & Grandvalet, 2003) technique for $k = 10$. The cross-validation method randomly divides the datasets into 10 equal partitions and uses 9 of them to train the model and the remaining one for testing the performance. The technique interchanges the training and testing set and repeats the process 10 times. In the testing phase, we utilize the best-performing model saved from the validation phase and apply it to the testing samples. Our performance metrics, which include accuracy, precision, recall, and F1 score, are computed in this phase. We report the final performance on the held-out test dataset. To empirically validate the results and other architectural specifications, we run the model 50 times and report average results across these repetitions with corresponding standard deviations. We also introduced multiple random initializations (Thimm & Fiesler, 1995) of the NN to ensure the agnosticism of the model to different initial conditions. The resulting metrics in Table 1 are summarized across distinct initializations. The joint training of all three modalities allows the sources to interact and helps modality-specific subnetworks to optimally converge by leveraging learning from the other subnetworks. Moreover, we implement a multimodal regularization technique for the completeness of the experiments. This technique aims to eliminate bias by maximizing the functional entropies (Gat et al., 2020). We designed our implementation based on the existing regularizer and utilities.

4 | EXPERIMENTAL RESULTS & DISCUSSION

We analyzed three datasets, COBRE (Mayer et al., 2013), fBIRN (Keator et al., 2016), and MPRC (Adhikari et al., 2019). The merged dataset consists of 437 subjects, with 275 healthy controls (HCs) and

TABLE 1 Performance comparison of our proposed model with several baselines (unimodal, bi-modal, and tri-modal) and the state-of-the-art models for a similar task.

Models	Modalities	Data	Accuracy	Precision	Recall	F1 score
Autoencoder	Unimodal	fMRI	0.811 ± 0.15	0.587 ± 0.11	0.451 ± 0.05	0.510 ± 0.11
Multilayer perceptron		sMRI	0.782 ± 0.13	0.439 ± 0.12	0.477 ± 0.04	0.443 ± 0.09
bi-LSTM with attention		SNPs	0.673 ± 0.15	0.338 ± 0.13	0.498 ± 0.04	0.403 ± 0.08
Mid fusion	Bi-modal	sMRI + fMRI	0.835 ± 0.14	0.577 ± 0.09	0.478 ± 0.03	0.523 ± 0.01
Mid fusion		SNPs + sMRI	0.784 ± 0.18	0.397 ± 0.08	0.455 ± 0.03	0.424 ± 0.15
Mid fusion		SNPs + fMRI	0.813 ± 0.14	0.501 ± 0.05	0.443 ± 0.03	0.470 ± 0.01
Early fusion	Multimodal	sMRI + fMRI + SNPs	0.701 ± 0.11	0.575 ± 0.12	0.460 ± 0.03	0.511 ± 0.02
Late fusion		sMRI + fMRI + SNPs	0.781 ± 0.08	0.615 ± 0.12	0.400 ± 0.04	0.485 ± 0.03
Mid-fusion (Rahaman et al., 2021)		sMRI + fMRI + SNPs	0.876 ± 0.06	0.640 ± 0.07	0.501 ± 0.04	0.562 ± 0.04
Mid-fusion with attention (Rahaman et al., 2023)		sMRI + fMRI + SNPs	0.921 ± 0.02	0.798 ± 0.07	0.522 ± 0.04	0.631 ± 0.05
Mid-fusion with mBAM (Rahaman et al., 2022)		sMRI + fMRI + SNPs	0.932 ± 0.04	0.903 ± 0.06	0.561 ± 0.05	0.692 ± 0.04
Spatio-modality mid fusion		sMRI + fMRI + SNPs	0.941 ± 0.05	0.829 ± 0.06	0.601 ± 0.06	0.697 ± 0.05

Note: The performance metrics are presented as (mean ± standard deviation) across 50 repetitive runs.

162 subjects with SZ. The performance of our proposed model is benchmarked against several other models, including unimodal and multimodal baselines using imaging-genetics datasets. The performance comparison is detailed in Table 1. In early fusion, input data from all modalities are merged at the first step, and the resultant vector is then processed through an NN. For late fusion, we use different networks for distinct modalities. Each network classifies a sample based on the data it receives, and a max voting scheme (Morvant et al., 2014) is used to determine the final label for each instance. Our proposed spatio-modality attentive fusion model outperforms the comparing methods (Table 1), achieving an accuracy of 94.1% in differentiating SZ. Other performance metrics are either superior or comparable to the benchmark models. Since our dataset is slightly imbalanced, we used the harmonic mean of precision and recall given by the F1 score (Goutte & Gaussier, 2005) for the classification performance evaluation. Our proposed model achieves an F1 of 0.697 with a reasonable balance between precision and recalls expected in a biological population. Substantial performance improvements are achieved when the BAM is used for fusion (as shown in the last two rows of Table 1), highlighting the utility of BAM in multimodal fusion. The performance of the model using only genetic data is consistently low. This suggests that genomic data is solely an insufficient descriptor of phenomena, that is, discriminating SZ. However, the noteworthy observation is the ability of the genomic data to complement other modalities in multimodal settings, especially when spatio-modality fusion is employed. This underscores the capacity of the fusion module to leverage the contributions from various modalities with greater precision. Our model can suppress the features from less informative sources while prioritizing those from more informative ones, a desirable characteristic to effectively learn the multimodal representation of input data.

4.1 | Reproducibility and reliability of the results

Reproducibility is a fundamental aspect that speaks to the reliability and validity of the computational findings (Adali & Calhoun, 2022; Klapwijk et al., 2021). The crucial medical research domain such as neuroimaging signifies the utility of replication most. It allows for the independent verification of results across analytical frameworks, enhancing the adaptability of vital neurobiological outcomes. In this study, we evaluate the reproducibility of the model's performance on the neuroimaging population by employing cross-validation strategies to assess the generalizability and stability of the results across different subsets of the data. We combine three datasets in this study COBRE, fBIRN, and MPRC, and preprocess them under the same imaging protocols. Rationally, we create three parts of the combined dataset and we check the reproducibility of each subpopulation separately. For recording performance on each dataset, we train the model on two other datasets. For instance, COBRE performance is computed using the COBRE sample as held-out test data, and the model is trained on only samples from fBIRN and MPRC. Additionally, these experiments aid in eradicating the acquisition bias and explore the model's efficacy on a smaller population. Furthermore, we generate two random splits of the data for the completeness of the experiments. It utilizes randomization training/validation and testing cohort selection. In the first random split (RDS 1), we select 40% of held-out data for testing and the remaining for training and validation. In the second one (RDS 2), we randomly select another 30% for the held-out and the remaining for training and validation. Table 2 shows the model performance on these diverse randomizations of the dataset to provide a comprehensive understanding of the efficiency of such a framework in disease prediction. We observe the performances are reasonably reproducible from different settings of input data. Three parts of the combined dataset yield comparable results with the final performance of the proposed framework. Two random splits also exhibit proportionate metrics validating the reliability of the model in the downstream task, which is classification in our case. However, with more data for training (RSD 2 in Table 2), the model achieves slightly better performance—standard in data-driven approaches. We also validate the extracted salient attributes of the data contributing to the SZ characterization across these splits. The spatio-modality attention-based feature analysis in the following subsection is carried out on the summarized set of features stable through independent segments of the dataset. To evaluate the impact of dilation on contextual learning, we conducted experiments using distinct dilatation d values. These also contribute to the verification of the model's robustness on specific choice configurations. Figure 3 presents the results for three performance metrics: accuracy, precision, and F1-score. The bar graph showing the evaluation metrics also includes the confidence interval. It indicates a lower and upper bound of the data that describes a range or a corridor in which 95% of the predictions would fall given the actual true value. We employed dilated convolution to extract contextual information. The dilation rate $d = 1$

TABLE 2 Reproducible performance of the proposed method for diverse subsets of the data.

Dataset/Split	Training and validation	Held-out test data	Accuracy	Precision	Recall	F1-score
COBRE	fBIRN + MPRC	COBRE	0.920 ± 0.08	0.802 ± 0.07	0.575 ± 0.05	0.68 ± 0.06
fBIRN	COBRE + MPRC	fBIRN	0.931 ± 0.09	0.801 ± 0.06	0.586 ± 0.05	0.67 ± 0.05
MPRC	COBRE + fBIRN	MPRC	0.901 ± 0.12	0.787 ± 0.09	0.589 ± 0.07	0.674 ± 0.07
RDS 1	60% of total data	40% of total data	0.932 ± 0.08	0.801 ± 0.01	0.608 ± 0.05	0.691 ± 0.03
RDS 2	70% of total data	30% of total data	0.937 ± 0.07	0.854 ± 0.03	0.583 ± 0.04	0.693 ± 0.03

Note: The first three rows represent the performance on the independent dataset and the bottom two rows stand for random splits.

represents the standard convolution, while *d* greater than 1 denotes the dilated version. The model demonstrates superior performance with a dilation value of 2 and experiences a performance drop for higher values. These findings suggest that dilation helps in extracting contextual information through a larger RF and is beneficial for accurate prediction. Nevertheless, because our multimodal fused tensor has limited dimensions, larger dilation might skip important transient patterns in the data by adding more holes in the RF. Due to the limited dimensions of the concatenated tensor, we are unable to explore the sensitivity beyond a dilation ratio of 3.

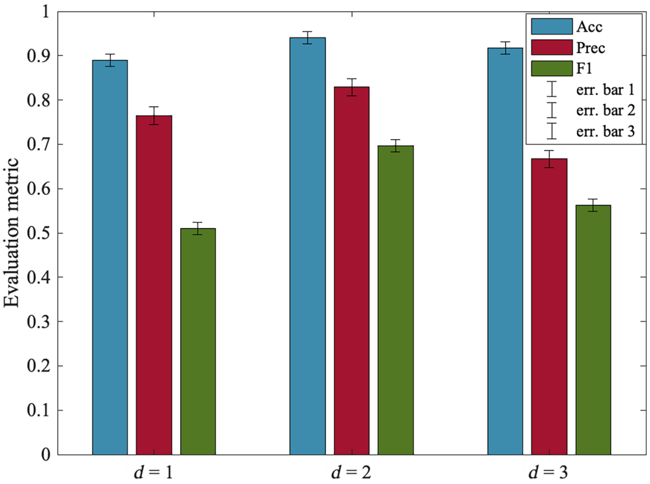

FIGURE 3 Model's performance for various dilation rates (*d*). The error bars stand for the confidence interval for the metrics accuracy (ACC), precision (Prec), and F1-score. The dilation rate indicates the expansion of the convolution kernel. Optimal performance is achieved at $d = 2$. The $d = 1$ represents the standard convolution. As the dilation rate increases beyond 2, we can see the performance start to decline.

4.2 | Spatio-modality attention for features attribution

In this experiment, we explore the feature's (from all modalities) contribution toward the disease classification. The proposed spatio-modality attention inscribes these significance scores. Thus, for feature attribution, we refine the attention values. However, to summarize the scores, averaging over multiple dimensions might drastically suppress the data, hence weakly depicting the tangible statistics. Here, we compute spatio-modality attention scores on a trained model by constraining multiple criteria. The values are summarized across the subjects and input feature dimensions. Furthermore, we compute scores feature-wise from all three modalities and determine a threshold representing the global mean throughout all the attention masks. We select the significant contributions by passing attention scores greater than the threshold only. Then, we filter the influential features by selecting significant contributions across at least 10% of the total population. The attention from all the settings mentioned in the reproducibility section is scrutinized to find a stable set of features. Therefore, we determine a contributing feature by their significant contribution toward classification across a reasonable number of samples. We avoid the features that are highly contributing to a few samples but are inconsistent across the population. As these sources are unstable and spurious, they are not reliable for neurological interpretation. Figure 4a shows the sMRI feature analysis based on

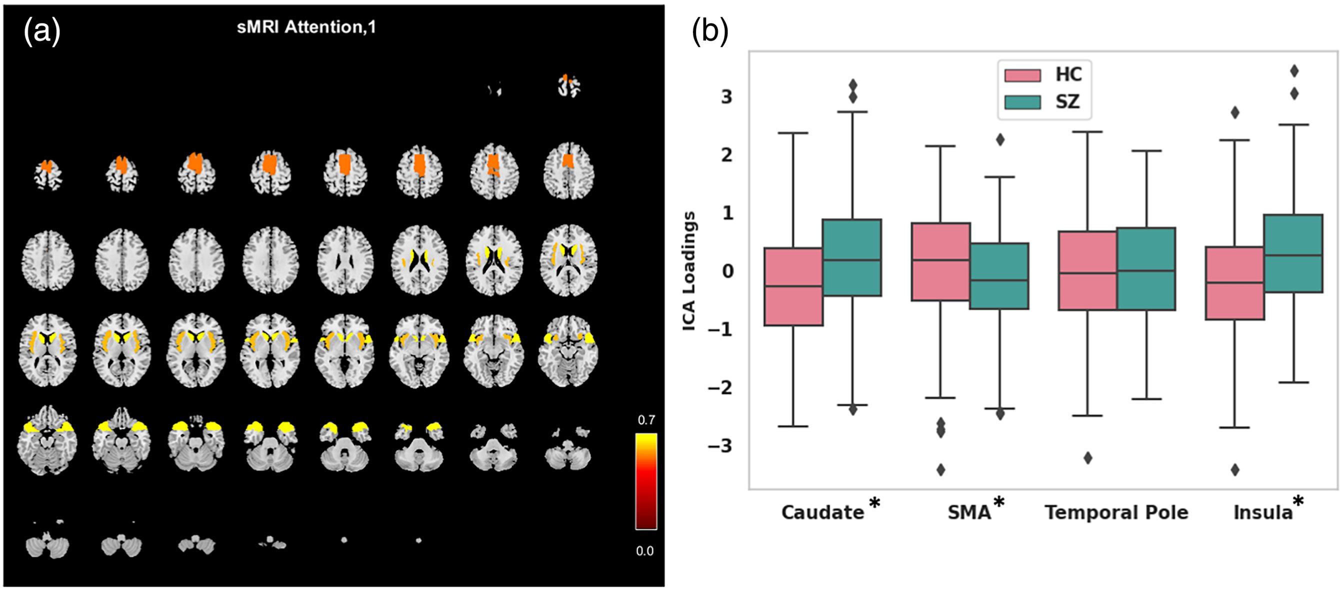

FIGURE 4 The spatio-modality attention-based sMRI features analysis. (a) Attention scores are computed on all sMRI components, and the significant ones are visualized on a structural montage. The components are supplementary motor area (SMA), caudate, temporal pole, and insula. (b) The ICA loadings for the most contributing components in HC and SZ groups. The error bars represent the standard deviation (SD). The asterisk sign on component name stands for the statistical significance of the difference between HC and SZ. We run two-sample *t*-test to validate the differences. We use *p*-value <.05 to mark the significant differences.

spatio-modality attention values. By thresholding the attention score, we observe four components to be significantly contributing to differentiating SZ. These components are the caudate, SMA, temporal pole, and insula. These components are the caudate, SMA, temporal pole, and insula. Moreover, we examine their ICA loadings on both HC and SZ groups and visualize the group-level statistics in Figure 4b. The HC loadings group mean is higher than the SZ mean, which indicates that HC subjects have higher gray matter density in the SMA region than SZ subjects. Alternatively, the SZ subjects are more heavily expressed in the insula and caudate than the HC group. Of note, the structural components differentiated through our method are associated with SZ, for example, caudate (Crespo-Facorro et al., 2007; Mueller et al., 2015) and temporal pole (John, 2009). Prior research also reported that aberrant motor behavior in SZ is associated with SMA volume (Schröder et al., 1995; Stegmayer et al., 2014). Also, insula activation has been associated with the processing of emotional facial expressions, which is deficient in individuals with SZ (Sheffield et al., 2020; Wylie & Tregellas, 2010). The differences in gray matter density patterns between these groups aid in understanding the brain's structural alterations associated with SZ. We also run the two-sample *t*-test on the loading values from both subject groups (HC/SZ) to verify the statistical significance of the group differences. We found three of the salient features (Caudate, SMA, and Insula) show statistically substantial differences between control and individuals with SZ at a level of $p < .05$. Furthermore, we analyze the attention computed from fMRI features. The static functional connections are shown in the connectograms of Figure 5. The connections are weighted by the attention scores in

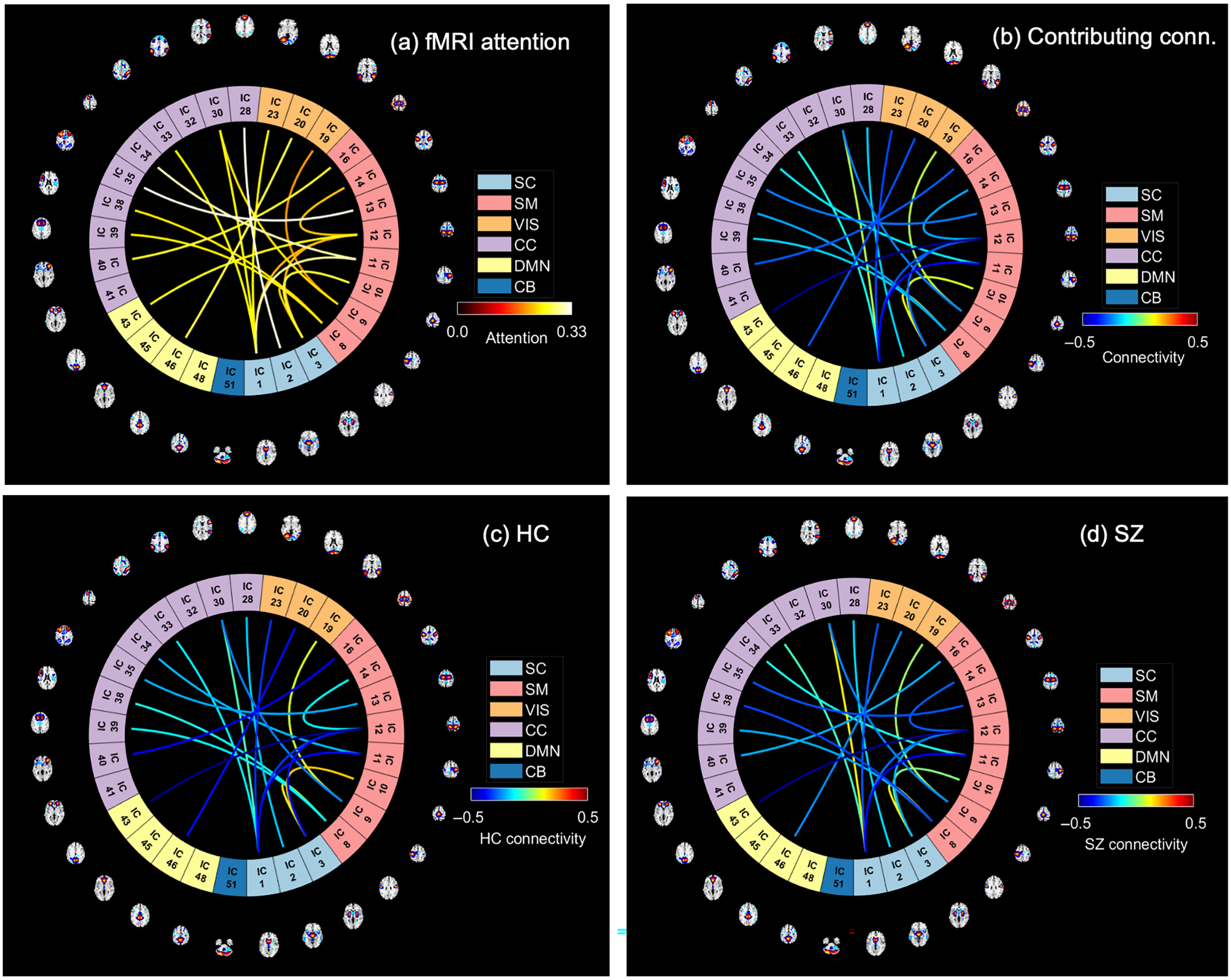

FIGURE 5 The spatio-modality attention scores computed from fMRI features (static functional connectivity (sFNC)). (a) The sFNC connections are weighted by their corresponding attention score. After thresholding, we show the connections that are effective for predicting schizophrenia at all different configurations of the model executions. The warm-colored connections are the significant ones; highly contributing to (a). We observe several contributing connections among sensorimotor (SM), cognitive control (CC), and subcortical (SC) regions. (b) The connections are weighted with the mean connectivity across the subject group. (c) Connections are weighted by the mean connectivity strength computed across the HC group. (d) SZ group's connectivity.

TABLE 3 The clusters extracted from biclustering on the attention values from all three modalities.

Bicluster	SZ subjects	HC subjects	sMRI features. (components loadings)	fMRI features (sFNC connections)	Genomics features (SNPs)
1	30	21	ACC + mpFC, Caudate, Temporal Pole, Frontal	# connections: 26	SNP IDs: rs10927041, rs10108628, rs10931784 Associated genes: AKT3, CSMD1 MOB4, HSPE1-MOB4
2	23	18	-	# connections: 73	-

Note: The N-BiC primarily extracts four biclusters that are merged into two due to higher overlaps with the earlier ones. Bicluster 1 is multimodal and bicluster 2 is unimodal consisting of 73 sFNC connections.
Abbreviation: ACC, anterior cingulate cortex.

Figure 5a. The connections among SM, CC, and SC are identified as most salient for classification. We further analyze the static functional connectivity of these edges between distinct brain components. Figure 5b shows average connectivity strengths are mostly negative (blue) and a few are positive (yellow). That demonstrates that most components are inversely correlated. The HC connectograms in Figure 5c show a connection with positive connectivity strength between IC 3 of SC to IC 10 of SM, whereas a positive connection in SZ dynamic is visible between IC 1 from SC to IC 30 of the CC domain. The HC dynamics are more strongly connected in the SC and SM regions than the dynamics of SZ subjects. The average functional connectivity strength between the components of these two regions is higher for HC subjects than for SZ. The visual (VS) domain also shows connectedness with other domains in HC dynamics where the SZ connections are comparatively weaker. The weaker connectivity in SZ might create functional impairments and dysfunction (Kaufmann et al., 2015; Koshiyama et al., 2018). Moreover, these connectivities also symbolize the communication between different parts of the brain for carrying out neural activities. Further study of these connections is required to assess their roles in overall cognition and information processing within these subject groups.

4.3 | Biclustering using spatio-modality attention scores

The spatio-modality attention masks provide the significance scores for the features from all modalities. In this experiment, we run biclustering of these scores across modalities to examine the subgrouping of subjects and features based on their association with SZ. Another rationale behind this experiment is to visualize the relation among distinct biological domains and their coregulation in diseased conditions. To run the analysis, first, we concatenate the feature's attention from sMRI, fMRI, and genomics modality, which is a two-dimensional matrix of (subject × features). Then, we run the N-BiC biclustering (Rahaman et al., 2020) for clustering the attention scores in both dimensions. We choose N-BiC because it allows clustering without specifying the expected number of biclusters and regulates the overlap between clusters. For two-dimensional data, it requires a heuristic to represent attributes from one variable as a function of attributes from another variable. It then exhaustively searches for all intrinsic subgroups and conditionally merges along the way. We sort features based on their consistently significant contributions to the classification. We select a subset of subjects for each feature where the attention values are higher than the global mean – the average attention across all the subjects and features. We choose features that exhibit this higher value across at least 10% of the total subjects. Initially, N-BiC yielded four biclusters then merged into two to regulate overlap. We repeat the run 10 times to stabilize the results. The outcomes show one multimodal and one unimodal bicluster (Table 3). The multimodal bicluster includes the following sMRI components: the anterior cingulate cortex, the medial prefrontal cortex, the caudate, the temporal pole, and the Frontal. It groups three SNPs including four genes CSMD1, ATK3, MOB4, HSPE1, and 26 static functional connections among several brain regions. AKT3 provides instructions for making a protein that is mostly active in the nervous system. The gene creates learning and memory-related deficits, and the SNP is identified as an associated genome-wide significant locus for SZ (Howell et al., 2017; Howell & Law, 2020). CSMD1 is known as a complement regulatory gene and has also been associated with SZ (Athanasiu et al., 2017; Liu et al., 2019). Figure 6 visualizes the sFNC connections through HC, SZ, and HC-SZ connectograms, respectively. In general, HC and SZ connections appear homogenous and rational since they are clustered into a single subgroup. However, one connection emerged with a strong HC-SZ difference between the CC (IC 28) and SM network (IC 12) even with the maximized homogeneity. These multimodal features analogously contributed to characterizing SZ; therefore, it potentially provides a link among these physiologies for disease-related deficits. These strong homogenous associations with SZ suggest potential co-aberration of these physiologies (genomics and brain) suggesting that further exploration of their coupling and progression may help improve our understanding of the disorder. Further analysis of these features would be productive to infer their joint modulation in SZ conditions. Also, a tri-modal feature set can potentially help explain the behavioral deficits from multiple biological perspectives.

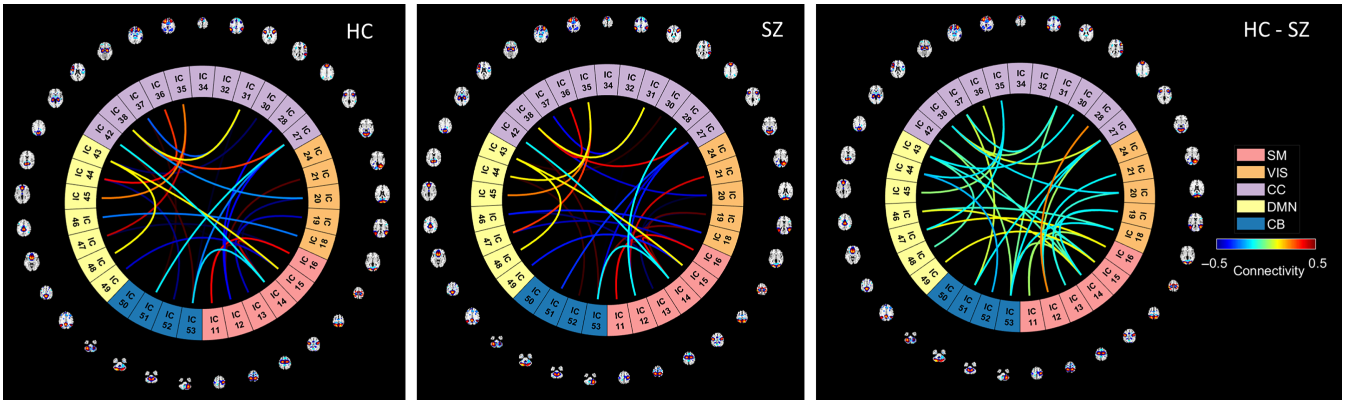

FIGURE 6 Connectograms generated from sFNC connections included in the multimodal bicluster (bicluster 1). There are HC/SZ group differences in the strength of static functional connectivity in cognitive control (CO) to sensorimotor (SM), and visual (VS) regions.

5 | CONCLUSION

Our proposed fusion model attentively probes the dimensions of the fused latent space and provides a more rational synthesis of multiple biological sources. The spatio-modality attention module boosts the classification performance and achieves an accuracy of 94.1% with a 0.697 F1 score. The noteworthy performance of the classifier with spatio-modality attentive fusion evidences the utility of such a technique in multimodal learning. Moreover, the spatio-modality attention scores are potentially self-explaining for the feature attributions toward the downstream task and provide an avenue for running further experiments, for example, subgrouping, and illustrating group differences. Our model learns spatial patterns through a large (dilated) RF for a better representation. The usage of dilated convolution for spatial scoring is potentially effective for foraging contextual information. In multimodal settings, the dilation is shown to be providing a comprehensive view of the available modes of data. The modality submodule signifies each source and generates modality-wise contributions. As such, the controlled scoring also interprets the model's decision and offers insights into the neurobiological relevance. This relevance can potentially explain the underlying substrate of the disorder under investigation, SZ. In all, the model seeks relevant patterns in the brain's structural features, functional mechanisms, and genomic pathways that lead to a coherent deciphering of SZ. Additionally, the analysis of modality-specific contributions helps discriminate SZ-affected physiologies with limited availability. The subgrouping of subjects and multimodal features based on the attention scores can potentially manifest the co-regulation of multiple biological domains in diseased conditions. The proposed fusion module can discover reliable biomarkers for the disorder, and the preceding interpretation recommends features that can help explain the underlying mechanism of the disease.

AUTHOR CONTRIBUTIONS

Md Abdur Rahaman and Yash Garg instantiated a bottleneck attention module for multimodal data fusion. Md Abdur Rahaman and Vince Calhoun designed the spatio-modality multisource fusion applied on neuroimaging genetics data. Md Abdur Rahaman ran the experiments and drafted the manuscript. Zening Fu, Armin Iraji, and Jiayu Chen helped to preprocess the datasets and edited the manuscript. Peter Kochunov, L. Elliot Hong, Theo G. M. Van Erp, and Adrian Preda collaborated in data collection and edited the manuscript. Vince Calhoun supervised all aspects of the project, thoroughly edited the paper, and gave valuable feedback on the results. All authors have approved the final version of the submission.

ACKNOWLEDGMENTS

The study was funded by National Institute of Health (NIH) (grant number R01MH118695) and National Science Foundation (NSF) (grant number 2112455).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

According to the IRB data privacy agreement, we are restricted to share any subject-specific data. However, the datasets are available online following the referred studies and also available from the corresponding authors on request. The preprocessing pipeline is public, as referred in the manuscript. All the preprocessing tools are available online at http://trendscenter.org/software, and the model's architectural details are available from the corresponding author.

ORCID

Md Abdur Rahaman https://orcid.org/0000-0002-4241-2439

REFERENCES

Abrol, A., Damaraju, E., Miller, R. L., Stephen, J. M., Claus, E. D., Mayer, A. R., & Calhoun, V. D. (2017). Replicability of time-varying connectivity patterns in large resting state fMRI samples. *NeuroImage*, *163*, 160–176.

Adali, T.l., & Calhoun, V. D. (2022). Reproducibility and replicability in neuroimaging data analysis. *Current Opinion in Neurology*, *35*(4), 475–481.

Adhikari, B. M., Hong, L. E., Sampath, H., Chiappelli, J., Jahanshad, N., Thompson, P. M., Rowland, L. M., Calhoun, V. D., du, X., Chen, S., &

Kochunov, P. (2019). Functional network connectivity impairments and core cognitive deficits in schizophrenia. *Human Brain Mapping*, *40*(16), 4593–4605.

Aine, C., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., Hanlon, F. M., Houck, J. M., Jung, R. E., Lauriello, J., & Liu, J. (2017). Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics*, *15*(4), 343–364.

Ashish, V. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Athanasiu, L., Giddaluru, S., Fernandes, C., Christoforou, A., Reinvang, I., Lundervold, A. J., Nilsson, L. G., Kauppi, K., Adolfsson, R., Eriksson, E., Sundet, K., Djurovic, S., Espeseth, T., Nyberg, L., Steen, V. M., Andreassen, O. A., & le Hellard, S. (2017). A genetic association study of CSMD1 and CSMD2 with cognitive function. *Brain, Behavior, and Immunity*, *61*, 209–216.

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2021). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer*, *37*, 1–32.

Bengio, Y., & Grandvalet, Y. (2003). No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, *16*, 1089–1105.

Bogdan, R., Salmeron, B. J., Carey, C. E., Agrawal, A., Calhoun, V. D., Garavan, H., Hariri, A. R., Heinz, A., Hill, M. N., Holmes, A., Kalin, N. H., & Goldman, D. (2017). Imaging genetics and genomics in psychiatry: A critical review of progress and potential. *Biological Psychiatry*, *82*(3), 165–175.

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6), 121.

Calhoun, V. D., & Allen, E. (2013). Extracting intrinsic functional networks with feature-based group independent component analysis. *Psychometrika*, *78*, 243–259.

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: Cognitive Neuroscience and Neuroimaging*, *1*(3), 230–244.

Chen, J., Calhoun, V. D., Pearlson, G. D., Perrone-Bizzozero, N., Sui, J., Turner, J. A., Bustillo, J. R., Ehrlich, S., Sponheim, S. R., Cañive, J. M., Ho, B. C., & Liu, J. (2013). Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *NeuroImage*, *83*, 384–396.

Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2017). Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, *7*(4), 750–758.

Crespo-Facorro, B., Roiz-Santiáñez, R., Pelayo-Terán, J. M., González-Blanch, C., Pérez-Iglesias, R., Gutiérrez, A., de Lucas, E. M., Tordesillas, D., & Vázquez-Barquero, J. L. (2007). Caudate nucleus volume and its clinical and cognitive correlations in first episode schizophrenia. *Schizophrenia Research*, *91*(1), 87–96.

Fu, Z., Iraji, A., Turner, J. A., Sui, J., Miller, R., Pearlson, G. D., & Calhoun, V. D. (2021). Dynamic state with covarying brain activity-connectivity: On the pathophysiology of schizophrenia. *NeuroImage*, *224*, 117385.

Fu, Z., Sui, J., Turner, J. A., du, Y., Assaf, M., Pearlson, G. D., & Calhoun, V. D. (2021). Dynamic functional network reconfiguration underlying the pathophysiology of schizophrenia and autism spectrum disorder. *Human Brain Mapping*, *42*(1), 80–94.

Gat, I., Schwartz, I., Schwing, A., & Hazan, T. (2020). Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. arXiv preprint arXiv:2010.10802.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European conference on information retrieval*. Springer.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, *7*, 63373–63394.

Gupta, C. N., Turner, J. A., & Calhoun, V. D. (2019). Source-based morphometry: A decade of covarying structural brain patterns. *Brain Structure and Function*, *224*(9), 3031–3044.

Hardoon, D. R., Ettinger, U., Mourão-Miranda, J., Antonova, E., Collier, D., Kumari, V., Williams, S. C. R., & Brammer, M. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neuroscience Letters*, *450*(3), 281–286.

Hazarika, D., Gorantla, S., Poria, S., & Zimmermann, R. (2018). Self-attentive feature-level fusion for multimodal emotion detection. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, Z. H., Hu, Y., Li, Y. C., Gong, L. J., Cieszczyk, P., Maciejewska-Karlowska, A., Leonska-Duniec, A., Muniesa, C. A., Marín-Peiro, M., Santiago, C., Garatachea, N., Eynon, N., & Lucia, A. (2015). PGC-related gene variants and elite endurance athletic status in a Chinese cohort: A functional study. *Scandinavian Journal of Medicine & Science in Sports*, *25*(2), 184–195.

Hong, C., Yu, J., Wan, J., Tao, D., & Wang, M. (2015). Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing*, *24*(12), 5659–5670.

Hosseinzadeh, M., & Wang, Y. (2021). Video captioning of future frames. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 980–989).

Howell, K. R., Floyd, K., & Law, A. J. (2017). PKBγ/AKT3 loss-of-function causes learning and memory deficits and deregulation of AKT/-mTORC2 signaling: Relevance for schizophrenia. *PLoS One*, *12*(5), e0175993.

Howell, K. R., & Law, A. J. (2020). Neurodevelopmental concepts of schizophrenia in the genome-wide association era: AKT/mTOR signaling as a pathological mediator of genetic and environmental programming during development. *Schizophrenia Research*, *217*, 95–104.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digital Medicine*, *3*(1), 1–9.

Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I., & Lungren, M. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: A case-study in pulmonary embolism detection. *Scientific Reports*, *10*(1), 1–9.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR.

Iraji, A., Faghiri, A., Fu, Z., Rachakonda, S., Kochunov, P., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., Mueller, B. A., Pearlson, G. D., Potkin, S. G., Preda, A., Turner, J. A., van Erp, T. G. M., & Calhoun, V. D. (2022). Multi-spatial-scale dynamic interactions between functional sources reveal sex-specific changes in schizophrenia. *Network Neuroscience*, *6*(2), 357–381.

Iraji, A., Faghiri, A., Lewis, N., Fu, Z., Rachakonda, S., & Calhoun, V. D. (2021). Tools of the trade: Estimating time-varying connectivity patterns from fMRI data. *Social Cognitive and Affective Neuroscience*, *16*(8), 849–874.

John, J. P. (2009). Fronto-temporal dysfunction in schizophrenia: A selective review. *Indian Journal of Psychiatry*, *51*(3), 180–190.

Kaufmann, T., Skåtun, K. C., Alnæs, D., Doan, N. T., Duff, E. P., Tønnesen, S., Roussos, E., Ueland, T., Aminoff, S. R., Lagerberg, T. V., Agartz, I., Melle, I. S., Smith, S. M., Andreassen, O. A., & Westlye, L. T.

(2015). Disintegration of sensorimotor brain networks in schizophrenia. *Schizophrenia Bulletin*, *41*(6), 1326–1335.

Keator, D. B., van Erp, T. G. M., Turner, J. A., Glover, G. H., Mueller, B. A., Liu, T. T., Voyvodic, J. T., Rasmussen, J., Calhoun, V. D., Lee, H. J., Toga, A. W., McEwen, S., Ford, J. M., Mathalon, D. H., Diaz, M., O'Leary, D. S., Jeremy Bockholt, H., Gadde, S., Preda, A., … Potkin, S. G. (2016). The function biomedical informatics research network data repository. *NeuroImage*, *124*, 1074–1079.

Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M., & Mills, K. L. (2021). Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience*, *47*, 100902.

Koshiyama, D., Fukunaga, M., Okada, N., Yamashita, F., Yamamori, H., Yasuda, Y., Fujimoto, M., Ohi, K., Fujino, H., Watanabe, Y., Kasai, K., & Hashimoto, R. (2018). Role of subcortical structures on cognitive and social function in schizophrenia. *Scientific Reports*, *8*(1), 1183.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2010). Multimodal images in the brain. In *The neurophysiological foundations of mental and motor imagery* (pp. 3–16). Oxford University Press.

Kumar, S. K. (2017). On weight initialization in deep neural networks. arXiv preprint arXiv:1704.08863.

Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., & Dehaene, S. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage*, *63*(1), 11–24.

Li, Z., Yang, F., & Luo, Y. (2019). Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation. *IEEE Access*, *7*, 72928–72935.

Liu, C., Sun, F., & Zhang, B. (2018). Brain-inspired multimodal learning based on neural networks. *Brain Science Advances*, *4*(1), 61–72.

Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, *8*, 29.

Liu, K., Li, Y., Xu, N., & Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730.

Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). Cptr: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804.

Liu, Y., Fu, X., Tang, Z., Li, C., Xu, Y., Zhang, F., Zhou, D., & Zhu, C. (2019). Altered expression of the CSMD1 gene in the peripheral blood of schizophrenia patients. *BMC Psychiatry*, *19*(1), 113.

MacDonald, A. W., III, Thermenos, H. W., Barch, D. M., & Seidman, L. J. (2009). Imaging genetic liability to schizophrenia: Systematic review of FMRI studies of patients' nonpsychotic relatives. *Schizophrenia Bulletin*, *35*(6), 1142–1162.

Mayer, A. R., Ruhl, D., Merideth, F., Ling, J., Hanlon, F. M., Bustillo, J., & Cañive, J. (2013). Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Human Brain Mapping*, *34*(9), 2302–2312.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning (CoNNL)*, Association for Computational Linguistics.

Menon, S. S., & Krishnamurthy, K. (2021). Multimodal ensemble deep learning to predict disruptive behavior disorders in children. *Frontiers in Neuroinformatics*, *15*, 15.

Meyer-Lindenberg, A. (2010). Imaging genetics of schizophrenia. *Dialogues in Clinical Neuroscience*, *12*(4), 449–456.

Morvant, E., Habrard, A., & Ayache, S. (2014). Majority vote of diverse classifiers for late fusion. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*. Springer.

Mueller, S., Wang, D., Pan, R., Holt, D. J., & Liu, H. (2015). Abnormalities in hemispheric specialization of caudate nucleus connectivity in schizophrenia. *JAMA Psychiatry*, *72*(6), 552–560.

Naderi, H., Soleimani, B. H., & Matwin, S. (2019). Multimodal deep learning for mental disorders prediction from audio speech samples. arXiv preprint arXiv:1909.01067.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, *34*, 11075–11084.

Nanay, B. (2018). Multimodal mental imagery. *Cortex*, *105*, 125–134.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *the International Conference on Machine Learning (ICML)* (pp. 689–696).

Palaskar, S., Sanabria, R., & Metze, F. (2018). End-to-end multimodal speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

Park, J., Woo, S., Lee, J. Y., & Kweon, I. S. (2018). BAM: Bottleneck attention module. arXiv preprint arXiv:1807.06514.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575.

Rahaman, M. A., Chen, J., Fu, Z., Lewis, N., Iraji, A., & Calhoun, V. D. (2021). Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. In *2021 43rd annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE.

Rahaman, M. A., Chen, J., Fu, Z., Lewis, N., Iraji, A., van Erp, T. G. M., & Calhoun, V. D. (2023). Deep multimodal predictome for studying mental disorders. *Human Brain Mapping*, *44*(2), 509–522.

Rahaman, M. A., Garg, Y., Iraji, A., Fu, Z., Chen, J., & Calhoun, V. (2022). Two-dimensional attentive fusion for multi-modal learning of neuroimaging and genomics data. In *2022 IEEE 32nd international workshop on machine learning for signal processing (MLSP)*. IEEE.

Rahaman, M. A., Mathalon, D., Lee, H. J., Jiang, W., Mueller, B. A., Andreassen, O., Agartz, I., Sponheim, S. R., Mayer, A. R., Stephen, J., Jung, R. E., Turner, J. A., Canive, J., Bustillo, J., Calhoun, V. D., Gupta, C. N., Rachakonda, S., Chen, J., Liu, J., … Ford, J. (2020). N-BiC: A method for multi-component and symptom biclustering of structural MRI data: Application to schizophrenia. *IEEE Transactions on Biomedical Engineering*, *67*(1), 110–121.

Roitberg, A., Pollert, T., Haurilet, M., Martin, M., & Stiefelhagen, R. (2019). Analysis of deep fusion strategies for multi-modal gesture recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, IEEE.

Saha, D. K., Silva, R. F., Baker, B. T., & Calhoun, V. D. (2022). Decentralized spatially constrained source-based morphometry. In *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*. IEEE.

Salgado-Pineda, P., Fakra, E., Delaveau, P., McKenna, P. J., Pomarol-Clotet, E., & Blin, O. (2011). Correlated structural and functional brain abnormalities in the default mode network in schizophrenia patients. *Schizophrenia Research*, *125*(2–3), 101–109.

Schröder, J., Wenz, F., Schad, L. R., Baudendistel, K., & Knopp, M. V. (1995). Sensorimotor cortex and supplementary motor area changes in schizophrenia: A study with functional magnetic resonance imaging. *The British Journal of Psychiatry*, *167*(2), 197–201.

Segall, J. M., Allen, E. A., Jung, R. E., Erhardt, E. B., Arja, S. K., Kiehl, K., & Calhoun, V. D. (2012). Correspondence between structure and function in the human brain at rest. *Frontiers in Neuroinformatics*, *6*, 10.

Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., Ravi, V., & Peters, A. (2020). A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, *194*, 105596.

Sheffield, J. M., Rogers, B. P., Blackford, J. U., Heckers, S., & Woodward, N. D. (2020). Insula functional connectivity in schizophrenia. *Schizophrenia Research*, *220*, 69–77.

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2017). Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, *22*(1), 173–183.

Sohn, K., Shang, W., & Lee, H. (2014). Improved multimodal deep learning with variation of information. *Advances in Neural Information Processing Systems*, *27*, 245–254.

Stegmayer, K., Horn, H., Federspiel, A., Razavi, N., Bracht, T., Laimböck, K., Strik, W., Dierks, T., Wiest, R., Müller, T. J., & Walther, S. (2014). Supplementary motor area (SMA) volume is associated with psychotic aberrant motor behaviour of patients with schizophrenia. *Psychiatry Research*, *223*(1), 49–51.

Summaira, J., Li, X., Shoib, A. M., Li, S., & Abdul, J. (2021). Recent advances and trends in multimodal deep learning: A review. arXiv preprint arXiv: 2105.11087.

Tang, S., Qi, Z., Granley, J., & Beyeler, M. (2021). U-net with hierarchical bottleneck attention for landmark detection in fundus images of the degenerated retina. In *Ophthalmic medical image analysis: 8th international workshop, OMIA 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, proceedings 8*. Springer.

Thimm, G., & Fiesler, E. (1995). Neural network initialization. In *From natural to artificial neural computation: International workshop on artificial neural networks Malaga-Torremolinos, Spain, June 7–9, 1995 proceedings 3*. Springer.

Tulay, E. E., Metin, B., Tarhan, N., & Arıkan, M. K. (2019). Multimodal neuroimaging: Basic concepts and classification of neuropsychiatric diseases. *Clinical EEG and Neuroscience*, *50*(1), 20–33.

Vaswani, K., Agrawal, Y., & Alluri, V. (2021). Multimodal fusion based attentive networks for sequential music recommendation. In *2021 IEEE seventh international conference on multimedia big data (BigMM)*. IEEE.

Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports*, *11*(1), 3254.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. Springer.

Wylie, K. P., & Tregellas, J. R. (2010). The role of the insula in schizophrenia. *Schizophrenia Research*, *123*(2–3), 93–104.

Xi, Y., Zhang, Y., Ding, S., & Wan, S. (2020). Visual question answering model based on visual relationship detection. *Signal Processing: Image Communication*, *80*, 115648.

Yaseen, M. U., Nasralla, M. M., Aslam, F., Ali, S. S., & Khattak, S. B. A. (2022). A novel approach based on multi-level bottleneck attention modules using self-guided dropblock for person re-identification. *IEEE Access*, *10*, 123160–123176.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yuhas, B. P., Goldstein, M. H., & Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, *27*(11), 65–71.

Zhang, W., Braden, B. B., Miranda, G., Shu, K., Wang, S., Liu, H., & Wang, Y. (2022). Integrating multimodal and longitudinal neuroimaging data with multi-source network representation learning. *Neuroinformatics*, *20*(2), 301–316.

How to cite this article: Rahaman, M. A., Garg, Y., Iraji, A., Fu, Z., Kochunov, P., Hong, L. E., Van Erp, T. G. M., Preda, A., Chen, J., & Calhoun, V. (2024). Imaging-genomic spatial-modality attentive fusion for studying neuropsychiatric disorders. *Human Brain Mapping*, *45*(17), e26799. https://doi.org/10.1002/hbm.26799