Multi-modal Spatial-modality Attentive Fusion for Studying Neuropsychiatric Disorders

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

[1] Georgia Institute of Technology, Atlanta, GA, USA, [2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Atlanta, GA, USA, [3] Nokia Bell Labs, NJ, USA, [4] University of Maryland Center for Brain Imaging Research, MD, USA, [5]Clinical Translational Neuroscience Laboratory, Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, California, USA, [6] Center for the Neurobiology of Learning and Memory, University of California Irvine, Irvine, California, USA, [7] Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, California, USA, [8]Department of Computer Science, Georgia State University, Atlanta, GA, USA

*Corresponding Author's Name: Md Abdur Rahaman, Email: mrahaman8@gatech.edu

***Abstract*—Multi-modal 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 domains-specific, leading to data being represented in dissimilar subspaces, and (ii) fused data is often noisy and high-dimensional, making it challenging to identify relevant information. To address these challenges, we propose a multi-modal artificial intelligence (AI) model with a novel fusion module inspired by a bottleneck attention module (BAM). We use deep neural networks (DNN) 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 multi-modal imaging-genetic dataset and achieve an SZ prediction accuracy of 94.10% (P < 0.0001), outperforming state-of-the-art unimodal and multi-modal 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 multi-modal cluster that includes genes CSMD1, ATK3, MOB4, and HSPE1, all of which have been identified as relevant to schizophrenia. 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.**

***Index Terms*—Multi-modal fusion, schizophrenia classification, bottleneck attention module (BAM), functional network connectivity (FNC), single nucleotide polymorphism (SNP), resting-state functional and structural MRI, imaging-genetics.**

I. INTRODUCTION

Our brain processes multi-modal signals from the outer world to understand real events and respond accordingly [1-5]. Leveraging information from diverse sources allows for a better understanding of a given phenomenon. Likewise, multi-modal learning enables researchers and practitioners to benefit from the unique strengths of each modality, leading to improved performance, enhanced accuracy, and richer insights [6, 7]. Applications of multi-modal learning models are wide-ranging and include fields such as computer vision [8-10], healthcare [11-13], and speech recognition [14, 15], among others [16]. Moreover, the potential applications of multi-modal learning in studying mental disorders are vast and continue to expand [17-22]. 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 [17, 22]. Genetic data provide another source of information regarding disease-related aberrations [23, 24]. 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 [22]. Particularly, neuroimaging modalities, when combined can achieve enhanced temporal and spatial resolution and bridge the gap between physiological and cognitive representations [25]. Hence, multi-modal learning frameworks have emerged as effective tools for analyzing data from multiple sources, including neuroimaging [17, 20, 26, 27] and genetic data [18, 28]. Past research has demonstrated significant correlations between structural and functional changes in the brain and mental disorders [29, 30]. Moreover, existing scientific literature points to a promising area of exploration: the

correlation between genetic variants and neural activity concerning neuropsychiatric disease-related degeneration [23, 24, 31]. Such a disorder Schizophrenia (SZ) is genetically complex and affects the brain's structure and function [18, 32, 33]. 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 is a layer with fewer neurons than the layer below or above it [34, 35]. 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 [36, 37]. 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 [36]. 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 neural networks are used to first learn the hidden space representations of the inputs. These networks translate the multi-modal inputs into a uniform embedding space, after which the mid-fusion module is used for the integration [37, 38]. 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 [39, 40], multiplicative combination layer [41], and transformer [42]. An early adaptation of the bottleneck attention module (BAM) [35] known as the mBAM [43] 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 [44] methods to learn the contextual pattern. The module explores spatial dimensions using a convolutional neural network (CNN) [45] augmented by a larger receptive field (RF), a feature known as dilated convolution. Using dilated convolutions significantly expands the receptive field, facilitating the collection of contextual patterns from the combined data [35]. 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 (Fig. 3). In earlier studies, the FC layers lost considerable spatial information while downsampling from the combined tensor to a one-dimensional vector [43]. 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 schizophrenia. Therefore, the mask can also be utilized to generate other analytics of the test samples e.g., 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 schizophrenia (162) and controls (275), intending to classify SZ. Our proposed method produces a multi-dimensional 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 sMRI components – such as the supplementary motor area (SMA), left insula, caudate, and temporal pole – of high importance for detecting schizophrenia (Figure 4). The attention scores on static functional connectivity suggest that several connections among the sensorimotor, subcortical, and cognitive control are particularly salient in schizophrenia. Biclustering the attention scores from three modalities discovers a multi-modal 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 schizophrenia. The primary features of our method are:

1. A fusion module capable of encoding both modality and spatial attention.

2. A dilated convolution to extract spatial patterns with a large receptive field.

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.

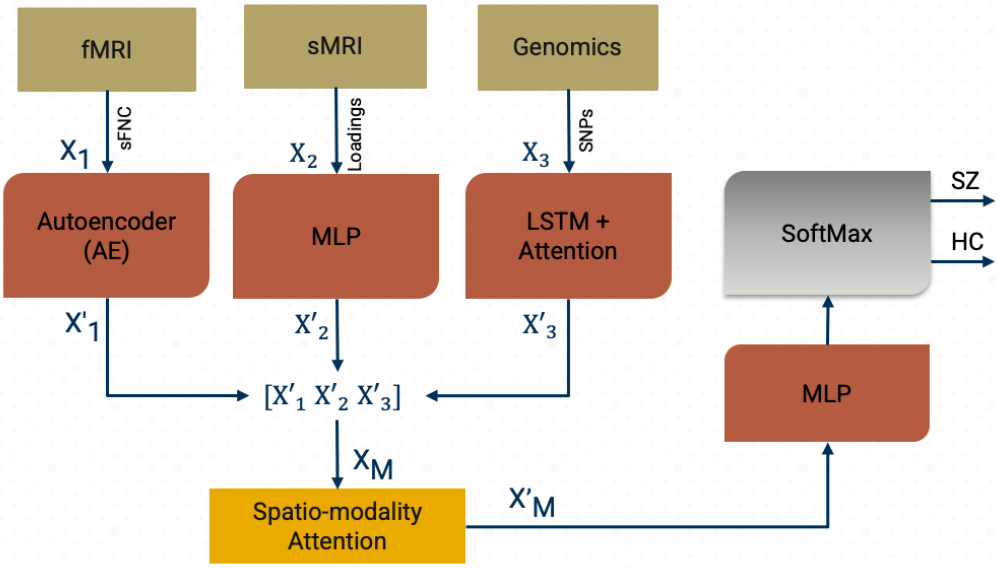

Fig 1. Our proposed multi-modal framework with spatio-modality temporal attention. These models incorporate three major processing steps: imaging-genomics preprocessing and dimensionality reduction, neural subnetworks for learning latent space embedding, and the predictor. We run group ICA on sMRI and fMRI data. We generate static functional network connectivity (sFNC) among the independent component networks (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 multi-layered 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.

II. Data Preprocessing

A. Structural MRI

We preprocessed structural magnetic resonance imaging (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 (GMV). Following this, we use a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm to smooth the gray matter densities.

B. 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 (MNI) standard space, using an echo-planar imaging (EPI) template, and resampled to $3 \times 3 \times 3$ mm^3 isotropic voxels. The images are then smoothed with a Gaussian kernel with FWHM = 6 mm, similar to the sMRI data.

C. Genomics

The preprocessing steps for the genetic data are described in our prior work [46, 47]. There are several standard preprocessing tools for genomics data, and we use Plink [48] for both preprocessing and imputation. Linkage disequilibrium (LD) pruning was administered at $r^2 < 0.9$. We use the Psychiatric Genomics Consortium (PGC) for SZ suggested GWAS [48] score to select the features. The analysis selects 1280 SNPs distributed across 108 risk loci. The PGC study [49] reveals these SNPs express statistically significant associations with schizophrenia at $p < 1\times10^{-4}$. 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 [50, 51].

III. Our Proposed Multi-modal Architecture

Figure 1 demonstrates our proposed architecture for multi-modal learning. The model incorporates three modality-specific neural networks (NN), referred to as subnetworks, which are used for learning the latent space embedding from each input source. These neural networks 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 multi-dimensional 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 multi-layer 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 neural networks to classify complex data effectively.

A. 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 Toolbox (GIFT) available at this link: (http://trendscenter.org/software/gift) [52] and the NeuroMark [46] template on the combined subjects from three datasets, which include FBIRN [53], COBRE [26], and MPRC [46]. The selection of independent 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 [50, 51]. The total number of connectivity networks extracted was 53, covering the whole brain. These intrinsic connectivity networks (ICNs) were distributed into seven functional domains [50, 54]: sub-cortical (SC), auditory (AD), sensorimotor (SM), visual (VS), cognitive-control (CC), default-mode (DM), and cerebellar domain (CB). 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 (TCs) 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) [55, 56], 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 [57]. 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 schizophrenia risk SNPs identified by the large PGC study.

B. Deep neural networks (DNNs)

We employ distinct 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) [58] 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 squared error (MSE). We use the sFNC matrix as an input from the fMRI modality and employ an AE with rectified linear unit (ReLU) 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 [59, 60]. We use Xavier initialization [61] 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 group ICA. The feature vector has a length of 30, where each value represents the expression level of a subject on a group structural independent component (GsIC). In general, loadings are just betas/coefficients of the variance mixture linear equations [62]. The loading value corresponds to how much variance a subject contributes to a given group component. Since the loadings are the vector of discrete values; for simplicity, we deployed a multi-layer perceptron (MLP) to efficiently learn the sMRI features. An MLP is a type of artificial neural network that consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. It can model complex, non-linear 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 schizophrenia. The subnetwork is fully connected 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 [63] to extract features from SNPs. The use of a bi-LSTM was informed by its anticipated ability to capture contextual information from a sequence [64, 65]. We assume that neighboring SNPs may show linkage disequilibrium 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 deep neural network's discriminative power.

C. Spatio-modality Attention

We propose a fusion module that takes a multi-modal 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 [35] which has been implemented for learning channels and spatial attention in classification and prediction tasks [66-68]. We adapt the module for fusing different modalities. The module is integrated with a mid/intermediate fusion [69] in a multi-modal

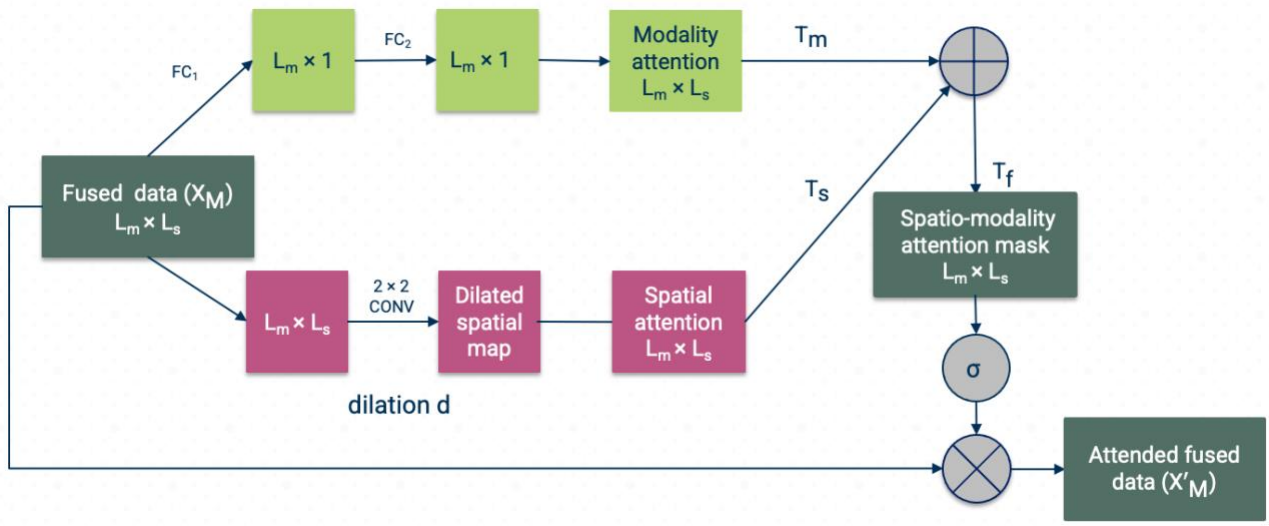

Fig. 2. Our proposed spatio-modality attention module for multi-modal fusion. The concatenated embeddings are sent through two different branches. i) The modality branch that learns the cross-modality interactions and mounts it into T_m attention mask, ii) the spatial branch (T_s) captures the relevant 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 multi-modal tensor and fully connected layer for modality attention.

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 fully connected (FC) linear layers in modality dimensions to learn the attention weights per modality. Fully Connected layer (FC), is a type of layer used in neural networks 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).

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

where $W \in \mathbb{R}^{N\times 1}$, $b \in \mathbb{R}^{N\times 1}$ is 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 fully connected layer.

For another branch, the module employs dilated convolution layers for extracting the relevant spatial patterns. The dilated convolutional filters arbitrate a large receptive field 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.

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 is then passed through another FC layer (FC_2) to compute the attention weights for each modality. We also use batch normalization (*BN*) [70] to adjust the scale with the spatial branch. Moreover, *BN* improves the speed, performance, and stability of the neural networks. Then, the tensor is expanded to the shape of the input tensor $L_m \times L_s$. We can formulate the operation as follows,

$$T_m = BN\ (\ FC_2\ (\ FC_1\ (X_M)\ )\ ) \tag{2}$$

Here, *Tm* refers to the transformed tensor after the application of batch normalization and two fully connected layers.

2) Spatial attention branch (Ts)

We opt for capturing the context from the aggregated multi-modal embedding tensor, which extracts contextual information in the form of spatial patterns. For spatial attention, the submodule uses a convolutional neural network. We use dilated convolution [44] to create a large receptive field that assists in apprehending the context. The dilated convolution inflates the kernel by inserting holes between the kernel elements [44]. 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 $L_m \times L_s$. Equation 3 is the operation for spatial attention and illustrates the overall processing in the spatial branch.

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

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
Multi-layer 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	Multi-modal	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 [18]		sMRI + fMRI + SNPs	0.876 ± 0.06	0.640 ± 0.07	0.501 ± 0.04	0.562 ± 0.04
Mid-fusion with attention [19]		sMRI + fMRI + SNPs	0.921 ± 0.02	0.798 ± 0.07	0.522 ± 0.04	0.631 ± 0.05
Mid-fusion with mBAM [43]		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

$$T_s = BN\,(\,d^{2\times2}(d^{2\times2}(X_M)) \qquad (3)$$

Here, T_s represents the spatial attention mask. The operation helps to create a large receptive field 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)).

$$T_f = \sigma\,(T_s + T_m\,) \qquad (4)$$

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 to 1. The sigmoid function is commonly used in neural networks to introduce non-linearity 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)).

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

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.

D. Multi-modal joint training

We split the dataset into two primary categories: 80% for training and 20% for evaluation. The evaluation data is 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 (LR) of 10^{-4} and a batch size of 32. The loss terms include mean square error (MSE) for the autoencoder 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 k-fold cross validation [71] 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 ten 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 deviation. We also introduced multiple random initializations [72] of the neural network to ensure the agnosticism of the model to different initial conditions. The resulting metrics in Table 1 are summarized across distinct intializations. 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 multi-modal regularization technique for the completeness of the experiments. This technique aims to eliminate bias by maximizing the functional entropies [73]. We designed our implementation based on the existing regularizer and utilities.

IV. Experimental Results & Discussion

We analyzed three datasets, COBRE [74], fBIRN [53], and MPRC [46]. The merged dataset consists of 437 subjects, with 275 healthy controls (HC) and 162 subjects with schizophrenia (SZ). The performance of our proposed model is benchmarked against several other models, including unimodal and multi-modal 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 a neural network. 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 [75] 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 schizophrenia. 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 [76] 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 bottleneck attentive module (BAM) is used for fusion (as shown in the last two rows of Table 1), highlighting the utility of BAM in multi-modal 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 i.e., discriminating schizophrenia. However, the noteworthy observation is the ability of the genomic data to complement other modalities in multi-modal 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 multi-modal representation of input data.

A. Reproducibility and reliability of the results

Reproducibility is a fundamental aspect that speaks to the reliability and validity of the computational findings [77, 78]. 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 preproces 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

Table. 2. Reproducible performance of the proposed method for diverse subsets of the data. The first three rows represent the performance on the independent dataset and the bottom two rows stand for random splits.

Dataset/Split	Training & 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

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 schizophrenia characterization across these

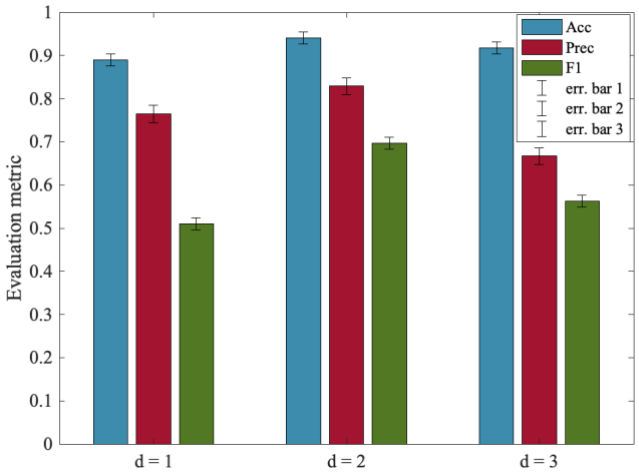

Fig. 3. Model's performance for various dilation rates (d). The error bars stand for the confidence interval for the metrics. The dilation rate indicates the expansion of the convolution kernel. Optimal performance is achieved at d = 2. The d = 1 represents the standard convolution. As the dilation rate increases beyond 2, we can see the performance start to decline.

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 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 receptive field and is beneficial for accurate prediction. Nevertheless, because our multi-modal fused tensor has limited dimensions, larger dilation might skip important transient patterns in the data by adding more holes in the receptive field. Due to the limited dimensions of the concatenated tensor, we are unable to explore the sensitivity beyond a dilation ratio of 3.

B. 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

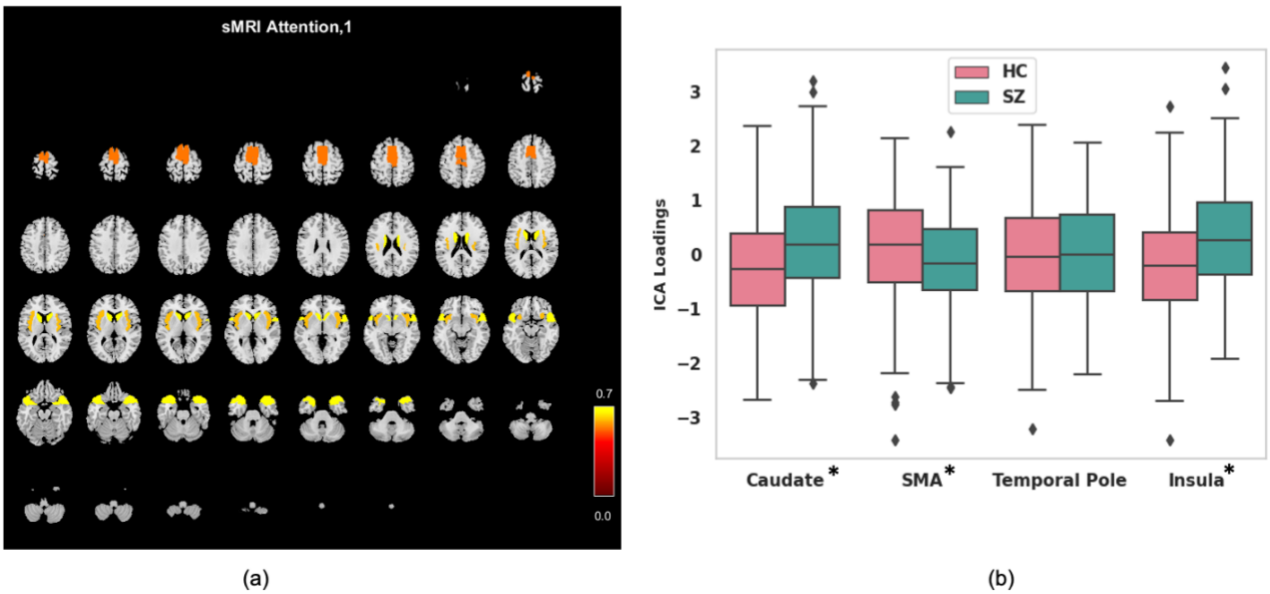

Fig. 4. The spatio-modality attention-based sMRI features analysis. (a) Attention scores are computed on all sMRI components, and the significant ones are visualized on a structural montage. The components are supplementary motor area (SMA), caudate, temporal pole, and insula. (b) The ICA loadings for the most contributing components in HC and SZ groups. The error bars represent the standard deviation (SD). The asterisks sign on component name stand for statistical significance of the difference between HC and SZ. We run two-sample t-test to validate the differences. We use p-value < 0.05 to mark the significant differences.

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 4 (a) shows the sMRI feature analysis based on spatio-modality attention values. By visualize the group-level statistics in Figure 4(b). 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 schizophrenia e.g., caudate [79, 80] and temporal pole [81]. Prior research also reported that aberrant motor behavior in schizophrenia is associated with supplementary motor area (SMA) volume [82, 83]. Also, insula activation has been associated with the processing of emotional facial expressions, which is deficient

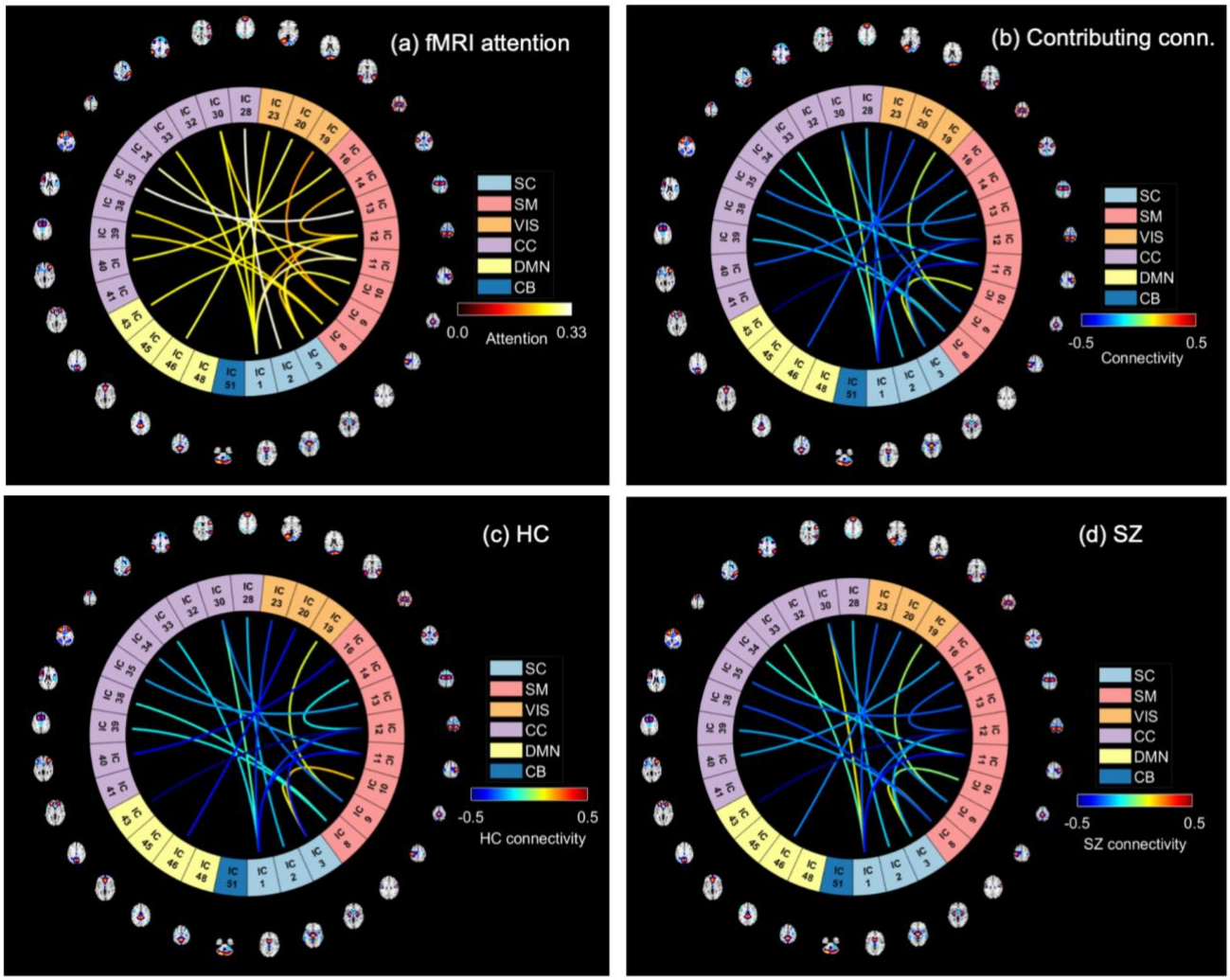

Fig. 5. The spatio-modality attention scores computed from fMRI features (static functional connectivity (sFNC)). (a) The sFNC connections are weighted by their corresponding attention score. After thresholding, we show the connections that are effective for predicting schizophrenia at all different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

thresholding the attention score, we observe four components to be significantly contributing to differentiating schizophrenia. These components are the caudate, supplementary motor area (SMA), temporal pole, and insula. These components are the caudate, supplementary motor area (SMA), temporal pole, and insula. Moreover, we examine their ICA loadings on both healthy control (HC) and schizophrenia (SZ) groups and in individuals with SZ [84, 85]. The differences in gray matter density patterns between these groups aid in understanding the brain's structural alterations associated with schizophrenia. 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

Table. 3. The clusters extracted from biclustering on the attention values from all three modalities. The N-BiC primarily extracts 4 biclusters that are merged into 2 due to higher overlaps with the earlier ones. Bicluster 1 is multi-modal and bicluster 2 is unimodal consisting of 73 sFNC connections.

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	-

differences between control and individuals with schizophrenia at a level of $p < 0.05$. 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 (a). The connections among sensorimotor (SM), cognitive control (CC), and subcortical (SC) are identified as most salient for classification. We further analyze the static functional connectivity of these edges between distinct brain components. Figure 5(b) 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 5(c) show a connection with positive connectivity strength between IC 3 of subcortical to IC 10 of sensorimotor, whereas a positive connection in SZ dynamic is visible between IC 1 from SC to IC 30 of the cognitive control domain. The HC dynamics are more strongly connected in the subcortical and sensorimotor 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 (VIS) 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 [86, 87]. 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.

C. 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 clustering of these scores across modalities to examine the subgrouping of subjects and features based on their association with schizophrenia. Another rationale behind this experiment is to visualize the relation among distinct biological domains and their coregulation in diseased conditions. To run the analysis, firstly, 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 [88] 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

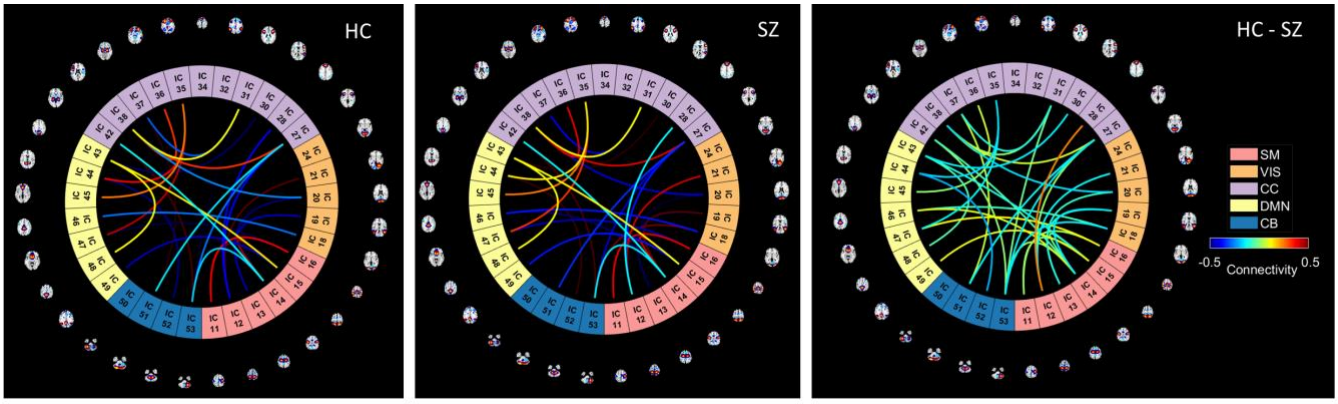

Fig. 6. Connectograms generated from sFNC connections included in the multi-modal bicluster (bicluster 1). There are HC/SZ group differences in the strength of static functional connectivity in cognitive control (CO) to sensorimotor (SM), and visual (VS) regions.

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 4 biclusters then merged into 2 to regulate overlap. We repeat the run 10 times to stabilize the results. The outcomes show 1 multi-modal and 1 unimodal bicluster (Table 3). The multi-modal bicluster includes the following sMRI components: the anterior cingulate cortex (ACC), 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 schizophrenia [89, 90]. CSMD1 is known as a complement regulatory gene and has also been associated with schizophrenia [91, 92]. 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 cognitive control (IC 28) and sensorimotor network (IC 12) even with the maximized homogeneity. These multi-modal features analogously contributed to characterizing SZ; therefore, it potentially provides a link among these physiologies for disease-related deficits. These strong homogenous associations with schizophrenia 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.

V. 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 multi-modal learning. Moreover, the spatio-modality attention scores are potentially self-explaining for the feature attributions towards the downstream task and provide an avenue for running further experiments e.g., subgrouping, and illustrating group differences. Our model learns spatial patterns through a large (dilated) receptive field for a better representation. The usage of dilated convolution for spatial scoring is potentially effective for foraging contextual information. In multi-modal 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 neurobiological relevance. This relevance can potentially explain the underlying substrate of the disorder under investigation, schizophrenia. 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 multi-modal 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.

Acknowledgments

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

Author Contributions

Md Abdur Rahaman (MAR) and Yash Garg (YG) initialized using a bottleneck attention module for multi-modal data fusion. MAR and Vince D. Calhoun (VDC) designed the spatio-modality multi-source fusion applied on imaging genetics data. MAR ran the experiments and drafted the manuscript. Zening Fu (ZU), Armin Iraji (AI), and Jiayu Chen (JC) helped to preprocess the datasets and edited the manuscript. Peter Kochunov, Elliot Hong, Theo G. M. Van Erp, and Adrian Preda collaborated in data collection and edited the manuscript.VDC 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.

Competing interests

There is no conflict of interest.

Code and Data availability statement

Based on the IRB data privacy agreement, we are not allowed to share any subject-specific data. However, the datasets are available online in the referred studies and can be requested from the corresponding authors suggested in the studies. The preprocessing pipelines are public, as referred to 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.

References

1. Bayoudh, K., et al., *A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets.* The Visual Computer, 2021: p. 1-32.
2. Guo, W., J. Wang, and S. Wang, *Deep multimodal representation learning: A survey.* IEEE Access, 2019. **7**: p. 63373-63394.
3. Summaira, J., et al., *Recent Advances and Trends in Multimodal Deep Learning: A Review.* arXiv preprint arXiv:2105.11087, 2021.
4. Kosslyn, S.M., G. Ganis, and W.L. Thompson, *Multimodal images in the brain.* The neurophysiological foundations of mental and motor imagery, 2010: p. 3-16.
5. Nanay, B., *Multimodal mental imagery.* Cortex, 2018. **105**: p. 125-134.

6. Ngiam, J., et al. *Multimodal deep learning*. in *ICML*. 2011.
7. Sohn, K., W. Shang, and H. Lee, *Improved multimodal deep learning with variation of information.* Advances in neural information processing systems, 2014. **27**.
8. Hosseinzadeh, M. and Y. Wang. *Video captioning of future frames*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.
9. Liu, W., et al., *Cptr: Full transformer network for image captioning.* arXiv preprint arXiv:2101.10804, 2021.
10. Xi, Y., et al., *Visual question answering model based on visual relationship detection.* Signal Processing: Image Communication, 2020. **80**: p. 115648.
11. Huang, S.-C., et al., *Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines.* NPJ digital medicine, 2020. **3**(1): p. 1-9.
12. Menon, S.S. and K. Krishnamurthy, *Multimodal Ensemble Deep Learning to Predict Disruptive Behavior Disorders in Children.* Frontiers in neuroinformatics, 2021. **15**.
13. Naderi, H., B.H. Soleimani, and S. Matwin, *Multimodal Deep Learning for Mental Disorders Prediction from Audio Speech Samples.* arXiv preprint arXiv:1909.01067, 2019.
14. Yuhas, B.P., M.H. Goldstein, and T.J. Sejnowski, *Integration of acoustic and visual speech signals using neural networks.* IEEE Communications Magazine, 1989. **27**(11): p. 65-71.
15. Palaskar, S., R. Sanabria, and F. Metze. *End-to-end multimodal speech recognition*. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. IEEE.
16. Sengupta, S., et al., *A review of deep learning with special emphasis on architectures, applications and recent trends.* Knowledge-Based Systems, 2020. **194**: p. 105596.
17. Calhoun, V.D. and J. Sui, *Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness.* Biological psychiatry: cognitive neuroscience and neuroimaging, 2016. **1**(3): p. 230-244.
18. Rahaman, M.A., et al. *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)*. 2021. IEEE.
19. Rahaman, M.A., et al., *Deep multimodal predictome for studying mental disorders.* Human Brain Mapping, 2023. **44**(2): p. 509-522.
20. Shi, J., et al., *Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease.* IEEE journal of biomedical and health informatics, 2017. **22**(1): p. 173-183.
21. Venugopalan, J., et al., *Multimodal deep learning models for early detection of Alzheimer's disease stage.* Scientific reports, 2021. **11**(1): p. 3254.
22. Zhang, W., et al., *Integrating multimodal and longitudinal neuroimaging data with multi-source network representation learning.* Neuroinformatics, 2022. **20**(2): p. 301-316.
23. Hardoon, D.R., et al., *Correlation-based multivariate analysis of genetic influence on brain volume.* Neuroscience letters, 2009. **450**(3): p. 281-286.
24. Le Floch, É., et al., *Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares.* Neuroimage, 2012. **63**(1): p. 11-24.
25. Liu, C., F. Sun, and B. Zhang, *Brain-inspired multimodal learning based on neural networks.* Brain Science Advances, 2018. **4**(1): p. 61-72.
26. Aine, C., et al., *Multimodal neuroimaging in schizophrenia: description and dissemination.* Neuroinformatics, 2017. **15**(4): p. 343-364.
27. Tulay, E.E., et al., *Multimodal neuroimaging: basic concepts and classification of neuropsychiatric diseases.* Clinical EEG and neuroscience, 2019. **50**(1): p. 20-33.
28. Bogdan, R., et al., *Imaging genetics and genomics in psychiatry: a critical review of progress and potential.* Biological psychiatry, 2017. **82**(3): p. 165-175.
29. Salgado-Pineda, P., et al., *Correlated structural and functional brain abnormalities in the default mode network in schizophrenia patients.* Schizophrenia research, 2011. **125**(2-3): p. 101-109.
30. Segall, J.M., et al., *Correspondence between structure and function in the human brain at rest.* Frontiers in neuroinformatics, 2012. **6**: p. 10.
31. Liu, J. and V.D. Calhoun, *A review of multivariate analyses in imaging genetics.* Frontiers in neuroinformatics, 2014. **8**: p. 29.
32. MacDonald III, A.W., et al., *Imaging genetic liability to schizophrenia: systematic review of FMRI studies of patients' nonpsychotic relatives.* Schizophrenia bulletin, 2009. **35**(6): p. 1142-1162.
33. Meyer-Lindenberg, A., *Imaging genetics of schizophrenia.* Dialogues in clinical neuroscience, 2010. **12**(4): p. 449.
34. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
35. Park, J., et al., *Bam: Bottleneck attention module.* arXiv preprint arXiv:1807.06514, 2018.
36. Huang, S.-C., et al., *Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection.* Scientific reports, 2020. **10**(1): p. 1-9.
37. Poria, S., et al., *A review of affective computing: From unimodal analysis to multimodal fusion.* Information fusion, 2017. **37**: p. 98-125.
38. Roitberg, A., et al. *Analysis of deep fusion strategies for multi-modal gesture recognition*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
39. Hazarika, D., et al. *Self-attentive feature-level fusion for multimodal emotion detection*. in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2018. IEEE.
40. Vaswani, K., Y. Agrawal, and V. Alluri. *Multimodal Fusion Based Attentive Networks for Sequential Music Recommendation*. in *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*. 2021. IEEE.
41. Liu, K., et al., *Learn to combine modalities in multimodal deep learning.* arXiv preprint arXiv:1805.11730, 2018.
42. Nagrani, A., et al., *Attention bottlenecks for multimodal fusion.* Advances in Neural Information Processing Systems, 2021. **34**.
43. Rahaman, M.A., et al. *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)*. 2022. IEEE.
44. Yu, F. and V. Koltun, *Multi-scale context aggregation by dilated convolutions.* arXiv preprint arXiv:1511.07122, 2015.
45. Gu, J., et al., *Recent advances in convolutional neural networks.* Pattern recognition, 2018. **77**: p. 354-377.
46. Adhikari, B.M., et al., *Functional network connectivity impairments and core cognitive deficits in schizophrenia.* Human brain mapping, 2019. **40**(16): p. 4593-4605.
47. Chen, J., et al., *Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference.* Neuroimage, 2013. **83**: p. 384-396.
48. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American journal of human genetics, 2007. **81**(3): p. 559-575.
49. He, Z.H., et al., *PGC-related gene variants and elite endurance athletic status in a C hinese cohort: A functional study.* Scandinavian journal of medicine & science in sports, 2015. **25**(2): p. 184-195.
50. Fu, Z., et al., *Dynamic state with covarying brain activity-connectivity: On the pathophysiology of schizophrenia.* Neuroimage, 2021. **224**: p. 117385.
51. Iraji, A., et al., *Multi-spatial-scale dynamic interactions between functional sources reveal sex-specific changes in schizophrenia.* Network Neuroscience, 2022. **6**(2): p. 357-381.
52. Iraji, A., et al., *Tools of the trade: estimating time-varying connectivity patterns from fMRI data.* Soc Cogn Affect Neurosci, 2021. **16**(8): p. 849-874.
53. Keator, D.B., et al., *The function biomedical informatics research network data repository.* Neuroimage, 2016. **124**: p. 1074-1079.
54. Fu, Z., et al., *Dynamic functional network reconfiguration underlying the pathophysiology of schizophrenia and autism spectrum disorder.* Human brain mapping, 2021. **42**(1): p. 80-94.
55. Gupta, C.N., J.A. Turner, and V.D. Calhoun, *Source-based morphometry: a decade of covarying structural brain patterns.* Brain Structure and Function, 2019. **224**(9): p. 3031-3044.

56. Saha, D.K., et al. *Decentralized Spatially Constrained Source-Based Morphometry*. in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022. IEEE.
57. Abrol, A., et al., *Replicability of time-varying connectivity patterns in large resting state fMRI samples.* Neuroimage, 2017. **163**: p. 160-176.
58. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
59. Chen, M., et al., *Deep feature learning for medical image analysis with convolutional autoencoder neural network.* IEEE Transactions on Big Data, 2017. **7**(4): p. 750-758.
60. Hong, C., et al., *Multimodal deep autoencoder for human pose recovery.* IEEE transactions on image processing, 2015. **24**(12): p. 5659-5670.
61. Kumar, S.K., *On weight initialization in deep neural networks.* arXiv preprint arXiv:1704.08863, 2017.
62. Calhoun, V.D. and E. Allen, *Extracting intrinsic functional networks with feature-based group independent component analysis.* Psychometrika, 2013. **78**: p. 243-259.
63. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
64. Li, Z., F. Yang, and Y. Luo, *Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation.* IEEE Access, 2019. **7**: p. 72928-72935.
65. Melamud, O., J. Goldberger, and I. Dagan. *context2vec: Learning generic context embedding with bidirectional lstm*. in *Proceedings of the 20th SIGNLL conference on computational natural language learning*. 2016.
66. Woo, S., et al. *Cbam: Convolutional block attention module*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.
67. Yaseen, M.U., et al., *A Novel Approach Based on Multi-Level Bottleneck Attention Modules Using Self-Guided Dropblock for Person Re-Identification.* Ieee Access, 2022. **10**: p. 123160-123176.
68. Tang, S., et al. *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*. 2021. Springer.
69. Boulahia, S.Y., et al., *Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition.* Machine Vision and Applications, 2021. **32**(6): p. 121.
70. Ioffe, S. and C. Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. in *International conference on machine learning*. 2015. PMLR.
71. Bengio, Y. and Y. Grandvalet, *No unbiased estimator of the variance of k-fold cross-validation.* Advances in Neural Information Processing Systems, 2003. **16**.
72. Thimm, G. and E. Fiesler. *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*. 1995. Springer.
73. Gat, I., et al., *Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies.* arXiv preprint arXiv:2010.10802, 2020.
74. Mayer, A.R., et al., *Functional imaging of the hemodynamic sensory gating response in schizophrenia.* Human brain mapping, 2013. **34**(9): p. 2302-2312.
75. Morvant, E., A. Habrard, and S. Ayache. *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)*. 2014. Springer.
76. Goutte, C. and E. Gaussier. *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation*. in *European conference on information retrieval*. 2005. Springer.
77. Adali, T.l. and V.D. Calhoun, *Reproducibility and replicability in neuroimaging data analysis.* Current Opinion in Neurology, 2022. **35**(4): p. 475-481.
78. Klapwijk, E.T., et al., *Opportunities for increased reproducibility and replicability of developmental neuroimaging.* Developmental Cognitive Neuroscience, 2021. **47**: p. 100902.
79. Crespo-Facorro, B., et al., *Caudate nucleus volume and its clinical and cognitive correlations in first episode schizophrenia.* Schizophrenia Research, 2007. **91**(1): p. 87-96.
80. Mueller, S., et al., *Abnormalities in hemispheric specialization of caudate nucleus connectivity in schizophrenia.* JAMA Psychiatry, 2015. **72**(6): p. 552-60.
81. John, J.P., *Fronto-temporal dysfunction in schizophrenia: A selective review.* Indian J Psychiatry, 2009. **51**(3): p. 180-90.
82. Stegmayer, K., et al., *Supplementary motor area (SMA) volume is associated with psychotic aberrant motor behaviour of patients with schizophrenia.* Psychiatry Res, 2014. **223**(1): p. 49-51.
83. Schröder, J., et al., *Sensorimotor Cortex and Supplementary Motor Area Changes in Schizophrenia: A Study with Functional Magnetic Resonance Imaging.* The British Journal of Psychiatry, 1995. **167**(2): p. 197-201.
84. Wylie, K.P. and J.R. Tregellas, *The role of the insula in schizophrenia.* Schizophr Res, 2010. **123**(2-3): p. 93-104.
85. Sheffield, J.M., et al., *Insula functional connectivity in schizophrenia.* Schizophrenia Research, 2020. **220**: p. 69-77.
86. Koshiyama, D., et al., *Role of subcortical structures on cognitive and social function in schizophrenia.* Scientific Reports, 2018. **8**(1): p. 1183.
87. Kaufmann, T., et al., *Disintegration of Sensorimotor Brain Networks in Schizophrenia.* Schizophr Bull, 2015. **41**(6): p. 1326-35.
88. Rahaman, M.A., et al., *N-BiC: A Method for Multi-Component and Symptom Biclustering of Structural MRI Data: Application to Schizophrenia.* IEEE Trans Biomed Eng, 2020. **67**(1): p. 110-121.
89. Howell, K.R., K. Floyd, and A.J. Law, *PKBγ/AKT3 loss-of-function causes learning and memory deficits and deregulation of AKT/mTORC2 signaling: Relevance for schizophrenia.* PloS one, 2017. **12**(5): p. e0175993.
90. Howell, K.R. and A.J. Law, *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, 2020. **217**: p. 95-104.
91. Athanasiu, L., et al., *A genetic association study of CSMD1 and CSMD2 with cognitive function.* Brain, Behavior, and Immunity, 2017. **61**: p. 209-216.
92. Liu, Y., et al., *Altered expression of the CSMD1 gene in the peripheral blood of schizophrenia patients.* BMC Psychiatry, 2019. **19**(1): p. 113.