

# A NOVEL DEEP SUBSPACE LEARNING FRAMEWORK TO AUTOMATICALLY UNCOVER ASSESSMENT-SPECIFIC INDEPENDENT BRAIN NETWORKS

*Ishaan Batta, Anees Abrol, Vince Calhoun*

Center for Translational Research in Neuroimaging and Data Science (TReNDS):  
Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, USA

## ABSTRACT

We present a novel deep learning framework to automatically compute independently salient networks in the brain that characterize the underlying changes in the brain in association with clinically observed assessments. Unsupervised approaches for high-dimensional neuroimaging data focus on computing low-dimensional brain components for subsequent analysis, while supervised learning approaches aim for predictive performance and yielding a single list of associative feature importance, thus making it hard to interpret at the level of brain subsystems. Our approach integrates the goals of decomposition into lower dimensional subspaces and, identifying salient brain subsystems into a single automated framework. We first train a convolutional neural network on structural brain features to predict clinical assessments, followed by a multi-step decomposition in the saliency space to compute salient brain networks that intrinsically characterize the brain changes associated with the assessment. Through a repeated training procedure on an Alzheimer's disease (AD) dataset, we show that our method effectively computes AD-related salient brain subsystems directly from high-dimensional neuroimaging data, while maintaining predictive performance. Such approaches are crucial for data-driven biomarker development for brain disorders.

**Index Terms**— Deep Learning, Subspace Learning, Neuroimaging, Saliency Analysis, Alzheimer's Disease, Ageing, Independent Component Analysis

## 1. INTRODUCTION

Brain disorders often involve changes that affect structure as well as function of multiple subsystems of the brain. The structural differences that manifest in patients with increasing severity of neurodegenerative disorders like Alzheimer's disease involve an interplay of complex changes in the brain and often require nuanced frameworks to analyze. With the hope of developing relevant biomarkers for various brain disorders, neuroimaging techniques like magnetic resonance

imaging (MRI) have been very promising for looking into such changes in a considerably detailed manner [1]. While MRI produces a detailed map of the full brain with a millimeter resolution, analyzing the resultant high-dimensional data from cohorts to extract meaningful signatures for disorder-related changes has been challenging. Numerous statistical and machine learning (ML) approaches have focused on the goal of utilizing neuroimaging data to understand these associated brain changes across subjects as well as lifespan [2, 3].

A prominent way for such approaches to study brain changes in disorders is to identify the set of brain regions that are either associated with or are predictive of a particular disorder class and its clinical assessment scores [3]. Features for ML approaches created from Neuroimaging data are usually used to assess the importance of various brain regions and/or connections toward a predictive ML task. Standard ML approaches have often used various ways to handle the high dimensionality of neuroimaging data by reducing or decomposing it to create useful features [4]. While reducing it by averaging over brain voxels belonging to regions of interest (ROIs) has been commonly used [5], decomposition methods to produce data-driven features have also shown to be helpful for studying associations with various disorders [6, 7]. Many decomposition approaches have been developed using methods like principal component analysis (PCA) as well as independent component analysis (ICA) [6] to create data-driven brain components corresponding to areas in the brain and summarizing the high dimensional neuroimaging data into low dimensional representations. Such approaches have also found utility in combining information from multiple datasets [8] and handling multiple data modalities involving structural and functional features, while yielding subject-specific features for associative analysis [9].

Since the oncoming of deep learning (DL) models, the need for having to reduce high-dimensional data into a summarized set of low-dimensional features has been removed as an additional step for prediction and associative purposes. This is because DL models have been shown to encode robust representations directly from voxel-level data as compared to summarized features from standard ML models, while also yielding better prediction performance [10, 11].

Corresponding author: Ishaan Batta (ibatta@gsu.edu)

However, post hoc feature analysis on both standard ML and DL models yields a single set of region-level or voxel-level results that have to be further manually studied or assigned to multiple brain sub-systems that are affected by the disorder [12]. Given that brain disorders often have multiple overlapping variants [13], each affecting multiple sub-systems in the brain, thus demanding the need for analysis frameworks to create multi-set or multi-level summaries of involved complex changes in the brain [14, 15]. While decomposition methods like independent subspace analysis have been built upon such promise [6], they have been developed for only unsupervised decomposition on neuroimaging data, not involving any particular disorders, thus requiring further development of complicated posthoc statistical or manual methods to utilize such subspaces for finding disease-specific changes in the brain. To uncover such associated multi-level changes that are involved in a brain disorder, there is a need to develop novel frameworks that not only encode associative information in an effective manner, but are also able to automatically discern the multiple sets of brain sub-systems that are affected, instead of returning a large single set of features ranked according to their importance.

We present a new methodological approach aimed at utilizing high-dimensional neuroimaging data with deep learning, as well as identifying multiple subsystems in the brain that are associated with changes in target assessments for a given brain disorder. Our approach involves the use of a DL model based on a convolutional neural network to first learn predictive associations between voxel-level neuroimaging features and target clinical assessments, followed by active subspace learning and independent components analysis on the saliency space of the learned models to identify multiple important directions that characterize the change in the target clinical assessment with respect to structural features in the brain. We also perform a robust repeated analysis to ensure the identification of consistent underlying independent active subspaces. By testing our model on an Alzheimer's disease dataset, we show that our framework is able to: (a) successfully automate the process of identifying multiple subspaces in the brain, (b) compute subspaces in a semi-supervised manner such that they capture important brain changes with respect to clinical assessments of a given disorder, (c) utilize voxel-level information by engaging DL architectures, and (d) retain comparable predictive performance.

## 2. METHODS

### 2.1. Dataset and Pre-Processing

We used structural MRI data from the ADNI dataset (adni.loni.usc.edu), including only the first visit scans for 1733 subjects (950/783 M/F, age  $75.54 \pm 7.29$ ) with 468 controls (CN), 933 subjects with mild cognitive impairment (MCI) and 332 subjects with Alzheimer's Disease (AD). Pre-

processing was done using a standard SPM12 pipeline as in previous studies [11]. For feeding into the deep learning architecture, all maps were warped to the standard MNI space with dimensions  $121 \times 145 \times 121$ , followed by Gaussian smoothing (FWHM = 12 mm). As target variable for the analysis, we used mini-mental score examination (MMSE) score (mean 26.99, std 2.89, range 9 – 30), which is a global clinical assessment (at a scale of 0 – 30) of cognitive status used towards diagnosis of cognitive impairment and Alzheimer's disease.

### 2.2. Multimodal Deep CNN Classifier

The features obtained after the pre-processing step mentioned in subsection 2.1 above were input into a 3D variant of the AlexNet architecture [16] as shown in step 1 of ???. AlexNet has been shown to encode predictive features using voxel-level data successfully [10].

Training of the architecture on for MMSE score regression was done independently with stratified 10-fold external cross-validation with 10% data for internal validation during training. After training, the gradients computed with back-propagation were used for the subsequent subspace learning analysis on the mutually exclusive test sets spanning the whole dataset from the 10 folds of the external cross-validation procedure.

### 2.3. Active Subspace Learning (PCA step)

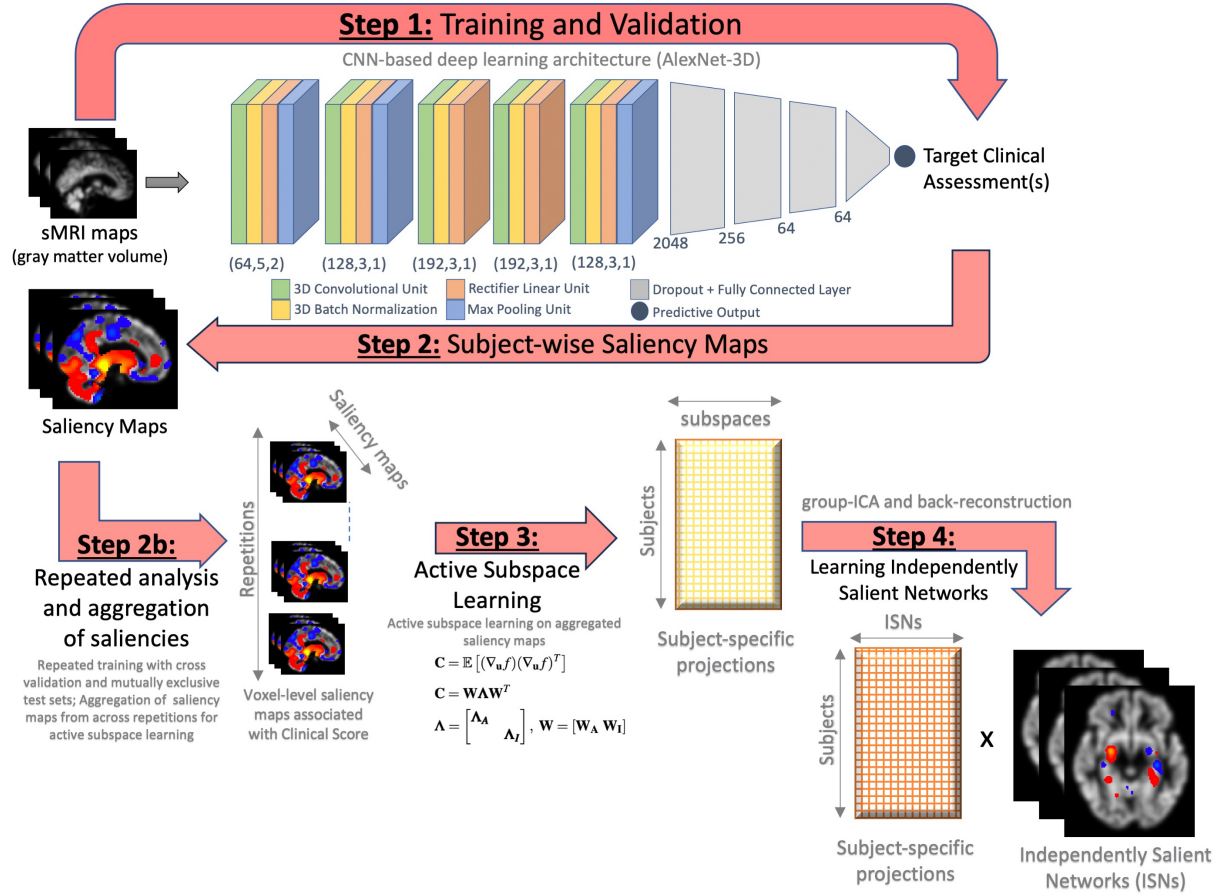
Let  $\mathbf{x} \in \mathbb{R}^m$  be a point in the  $m$ -dimensional space of input features, and consider a function  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  that maps input space to the target variable space. It should be noted that in this case,  $x$  could represent structural brain features like GMV map for a subject with  $y$  as the value of the clinically observed cognitive or biological assessment variable, and  $f$  can be the underlying regression function learned by the DL architecture upon training. Active subspace learning on Learning of active subspaces for the mapping  $f$  is performed as an eigendecomposition of covariance of the gradients of  $f$ . The covariance  $\mathbf{C}$  can be defined as:

$$\mathbf{C} = \mathbb{E} [(\nabla_{\mathbf{x}} f)(\nabla_{\mathbf{x}} f)^T] \quad (1)$$

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^n (\nabla f(\mathbf{x}_i))(\nabla f(\mathbf{x}_i))^T \quad (2)$$

In practice, one can estimate  $\mathbf{C}$  as  $\hat{\mathbf{C}}$  from the data. In this approach,  $f$  can be considered to be the function learned by the 3D-CNN architecture in Figure 1 upon being trained on a dataset with  $n$  subjects,  $[\mathbf{X}, \mathbf{y}]$ , where  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , and  $\mathbf{y} \in \mathbb{R}^n$ .

The next step in active subspace learning involves the eigendecomposition of  $\mathbf{C}$  to obtain a set of active subspaces represented by the eigenvectors with significantly large eigenvalues (Equation 4). Subsequently, transformed features  $\hat{\mathbf{X}}$



**Fig. 1:** An overview of the methodology to compute independently salient networks (ISNs) from structural MRI maps. After training a CNN-based DL architecture with repeated 10-fold cross-validation (step 1), the subject-specific saliency maps from across repetitions are aggregated (step 2) and decomposed (step 3) using an active subspace learning framework defined in subsection 2.3. Subsequently, ICA is performed (step 4) on the active subspaces followed by back-reconstruction to obtain the ISN maps.

can be generated by projecting the input data onto the active subspaces (Equation 5).

$$\mathbf{C} = \mathbf{W}\mathbf{A}\mathbf{W}^T \quad (3)$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_A & \mathbf{A}_I \end{bmatrix}, \mathbf{W} = [\mathbf{W}_A \ \mathbf{W}_I], \quad (4)$$

$$\text{such that } \mathbf{A}_I \approx \mathbf{0}, \text{ and } \lambda_i \gg 0 \ \forall \lambda_i \in \mathbf{A}_A$$

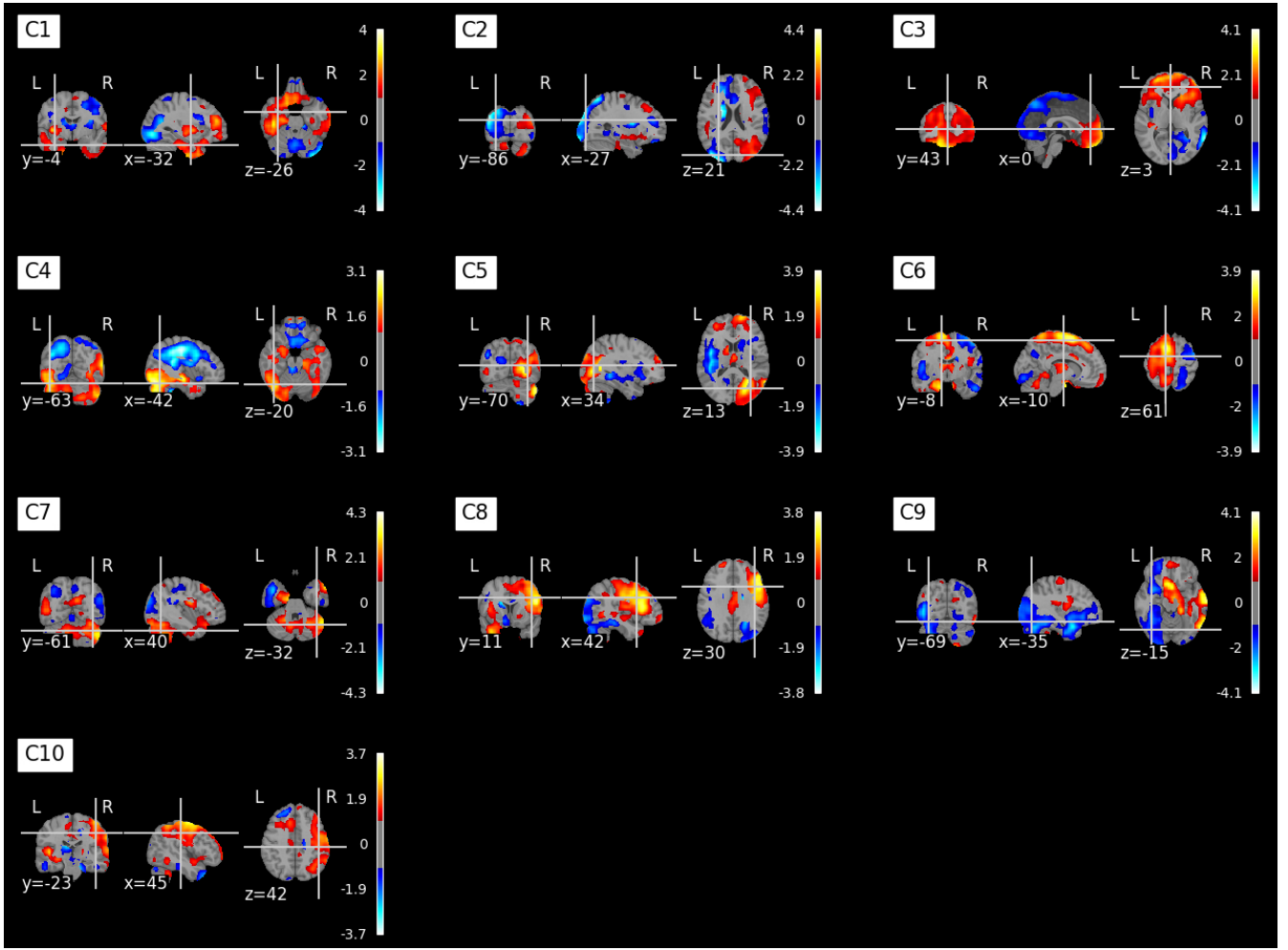
$$\hat{\mathbf{X}} = \mathbf{W}_A^T \mathbf{X} \quad (5)$$

#### 2.4. Learning of independently salient networks (ISNs)

The procedure of learning active subspaces from the data described in subsection 2.3 can be considered as performing principal component analysis (PCA) on the saliency space of the data since the active subspaces are essentially mutually orthogonal directions of maximum variance in the saliency. In the current context, the input feature space corresponds to the voxel-level structural MRI maps while the saliency space

corresponds to voxel-level saliency maps generated via gradient back-propagation from the trained CNN architecture as described in subsection 2.2. Thus, to compute the active subspaces in the data, we performed PCA on the saliency maps obtained from the test subjects across the 10-fold repetitions spanning the whole dataset. The model order was selected to get the top subspaces such that they cover at least 95% of the variance in the data.

For voxel-level neuroimaging data, independent component analysis (ICA) after the PCA step as dimensionality reduction is known to give more stable results in terms of computing meaningful sources in larger datasets because of it being a generative model assuming the sources in the data to be independent mixtures, which by definition also involves higher order statistics that just non-correlation maximized by PCA [6]. Thus, we performed an ICA on the subspace projections  $\hat{\mathbf{X}}$  computed by the PCA step on saliency maps from across repetitions, to obtain independently salient networks (ISNs) in the data. The ISNs characterize independent salient



**Fig. 2:** Standardized full-brain maps showing independently salient networks (ISNs) for MMSE score. ISNs are computed by performing a PCA step on the voxel-level saliency maps from a trained DL model (Figure 1) for MMSE regression using structural gray matter volume (GMV) maps, followed by ICA to compute the underlying directions of change in MMSE with respect to the changes in GMV in the brain.

directions of change in the target variables with respect to changes in the input. In addition to the ISNs, the ICA procedure also yields subject-specific loadings, which are the projection of input data onto the ISNs.

This ICA decomposition procedure, which acts on the PCA loadings matrix  $\hat{\mathbf{X}}$  from Equation 5, can be described as follows:

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{S} \quad (6)$$

$$\hat{\mathbf{S}} = \mathbf{W}_A \mathbf{S}^T \quad (7)$$

The matrix  $\mathbf{A}$  represents the ICA loadings and  $\mathbf{S}$  represents the ISNs computed as source ICA components but in terms of a weighted linear combination active subspaces in  $\mathbf{W}_A$ . The ISNs are computed as voxel-level brain maps by back-reconstruction as shown in Equation 7, representing

multiple independent intrinsically salient brain networks that characterize the changes in brain structure with respect to the changes in the clinically observed variable (MMSE).

### 3. RESULTS

#### 3.1. Computing Independently Salient Networks (ISNs)

Upon training the CNN model, the saliency maps were computed for test subjects from each repetition using gradient back-propagation followed by Gaussian smoothening (FWHM=12mm) and global mean removal. A PCA step was used to reduce the voxel-level saliency maps in MNI space to 1000 constituent active subspaces and loadings so as to cover at least 95% of variance in the saliency data. Subsequently, an ICA step with a model order of 10 was performed. This was followed by back-reconstruction as described in subsec-

ISN-1	ISN-2	ISN-3	ISN-4	ISN-5	ISN-6	ISN-7	ISN-8	ISN-9	ISN-10
RPall(1.73)	LAmyg(-1.94)	SpmGa(-1.35)	SMC(-2.54)	FrMC(-1.41)	SpmGa(1.26)	RAmyg(-2.36)	LPall(-1.98)	OccPole(1.16)	FrMC(-1.37)
RAmyg(1.39)	OccPole(-1.19)	LAmyg(1.26)	LPall(1.54)	CBv7-10(-1.34)	SpmGa(1.05)	LAmyg(-1.44)	LPut(-1.65)	SpmGa(-1.14)	LAmyg(1.10)
OccPole(-1.35)	LPall(-1.15)	OccPole(1.25)	SFG(-1.07)	LBrSt(-1.23)	AngG(0.97)	CBICr(0.91)	CBICr(-1.07)	SupCalc(1.08)	TOfusi(1.05)
LPut(1.27)	FrMC(0.98)	SPL(-1.13)	CBICr(-0.92)	LCaud(0.97)	LAmyg(-0.83)	CB11-4(0.84)	SpmGa(-0.85)	Cun(1.01)	LPut(1.00)
LAmyg(0.99)	RAmyg(-0.90)	SpmGa(-1.11)	LNA(-0.81)	LThal(0.88)	RPall(0.81)	CB17-10(0.83)	RAmyg(0.83)	iCalc(0.90)	FP(-0.97)
LHipp(0.97)	SFG(0.86)	PoCG(-1.06)	LOCs(0.80)	CeOper(0.87)	SFG(0.80)	CBv5-6(0.80)	MTGa(0.80)	SpmGa(-0.73)	Ins(0.70)
TOfusi(0.97)	CBr7-10(0.79)	AngG(-1.03)	STGp(0.79)	FrOrb(-0.82)	PoCG(0.75)	RCaud(0.78)	AngG(-0.80)	PC(0.71)	CB11-4(0.70)
LPall(0.94)	RCaud(0.75)	PreCG(-0.98)	Cun(0.77)	RThal(0.77)	FP(-0.75)	LHipp(0.77)	RThal(-0.78)	AngG(-0.70)	LPall(0.65)
ParOper(-0.93)	LOCi(-0.72)	SFG(-0.85)	LPut(0.74)	STGa(-0.75)	PreCG(0.73)	IFGpo(-0.76)	SpmGa(-0.77)	RPall(-0.63)	SFG(-0.64)
PlanTe(-0.90)	TP(0.70)	PC(-0.77)	MTGto(0.70)	CB11-4(-0.75)	SMC(0.72)	RHipp(-0.75)	TfusiA(-0.77)	RAmyg(-0.59)	CB15-6(0.63)
SC-VIS	SC-CC	DMN	CC-SM	SC-CC	SM-SC	DMN-SM	AUD-SC	VIS-DMN	CC-SC-AUD

**Table 1:** The top 10 regions of interest (ROIs) for all ISNs sorted based on mean contributive strength defined as mean of z-score of standardized ISN map across ROI voxels with  $|z| > 1$  in the given ISN map. Mean strengths are shown in parentheses with signs indicating whether the particular ROI had a positive or negative contribution in the ISN towards change in MMSE score. Full names and coordinates for the ROIs can be found at this link. The lowermost row for each ISN represents the primary brain domains of the involved ROIs based on their function. These domains are: the default mode network (DMN), visual areas (VIS), auditory areas (AU), cerebellar areas (CB), cognitive control (CC), sensorimotor (SM) and sub-cortical (SC) areas.

tion 2.4 to get brain maps for 10 independently salient networks (ISNs) that characterize the change in MMSE scores with respect to structural GMV features. Figure 2 shows the 10 back-reconstructed ISNs as standardized brain maps.

### 3.2. Prediction Performance

Hyperparameter tuning for batch size (bs) and learning rate (lr) was performed on the CNN architecture shown in Figure 1 to obtain an optimally performing model for MMSE regression (bs=32, lr=0.01). We did external 10-fold cross-validation, leading to 10 repetitions of the training and testing procedure on the dataset. Each repetition involved an internal validation on 10% samples taken out of training data for learning the optimal model. The model’s performance (Pearson correlation =  $0.56 \pm 0.07$ , MAE =  $1.83 \pm 0.06$ ) was comparable to previous works involving MMSE regression using deep learning on GMV maps [10]. It should be noted that in the scope of this study, the main aim was computing relevant ISNs, given the regression model learns with comparable performance to previous studies.

### 3.3. Biological relevance of ISNs

Table 1 shows the top 10 regions of interest (ROIs) in the brain and the involved brain domains from each of the ISN sorted according to the mean contributive strength across the ROI voxels calculated as the mean z-score across ROI voxels in the standardized ISN brain maps. As listed in Table 1, we found that the ISNs cover important regions in various brain domains namely the default mode network (DMN), visual areas (VIS), auditory areas (AU), cerebellar areas (CB), cognitive control (CC), sensorimotor (SM) and sub-cortical (SC) areas. For example, the first ISN corresponds to SC-VIS, the second one to SC-CC, the third to DMN, and so on. The set of domains and ROIs are interesting because many of these are the ones involved in what MMSE measures, namely:

orientation, attention, memory, language, and visual-spatial skills. Essentially, our framework is able to uncover independent brain subsystems that characterize the multiple aspects of changes captured by the composite MMSE score.

Brain areas like the hippocampus, amygdala, parahippocampal gyrus, occipital pole, fusiform gyrus, and cerebellum feature in the ISNs. Moreover, the thalamus, putamen, and caudate are also part of many ISNs. These observations agree with the results from earlier studies regarding the brain areas disrupted in Alzheimer’s disease [17].

## 4. CONCLUSION

In this work, we present a novel methodology to compute independent networks in the brain that characterize the saliency of brain structure towards changes captured by target clinical assessments (MMSE score) for a given disorder. Instead of performing unsupervised decompositions, our framework is able to take into account the saliency information from specific target clinical assessments while also utilizing voxel-level data in a deep learning framework. Moreover, instead of simply summarizing a list of important associated brain areas, our framework is aimed at an automated computation of intrinsic brain networks associated with a particular clinical variable. Our framework is able to successfully synthesize the goals of saliency analysis and subspace decomposition into a single automated pipeline while also handling voxel-level features from the brain.

In summary, such frameworks are essential for biomarker development for brain disorders as they offer an integrated approach for studying associations of the changes in the brain with the onset of brain disorders. In future, this approach can be extended to synthesize information from multimodal data as well as multiple clinical assessments at the same time.

## 5. ACKNOWLEDGMENTS

The work presented in this manuscript was supported financially by grants NIH RF1AG063153 and NSF 2112455 from the National Institutes of Health (NIH) and National Science Foundation (NSF) respectively.

## 6. REFERENCES

- [1] Avinash Chandra, George Dervenoulas, Marios Politis, and Alzheimer's Disease Neuroimaging Initiative, "Magnetic resonance imaging in alzheimer's disease and mild cognitive impairment," *Journal of neurology*, vol. 266, pp. 1293–1302, 2019.
- [2] Christos Davatzikos, "Machine learning in neuroimaging: Progress and challenges," *Neuroimage*, vol. 197, pp. 652, 2019.
- [3] Jing Sui, Rongtao Jiang, Juan Bustillo, and Vince Calhoun, "Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises," *Biological psychiatry*, vol. 88, no. 11, pp. 818–828, 2020.
- [4] Beatriz Remeseiro and Veronica Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in biology and medicine*, vol. 112, pp. 103375, 2019.
- [5] Saima Rathore, Mohamad Habes, Muhammad Ak-sam Iftikhar, Amanda Shacklett, and Christos Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, 2017.
- [6] Vince D Calhoun, Jingyu Liu, and Tülay Adalı, "A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data," *Neuroimage*, vol. 45, no. 1, pp. S163–S172, 2009.
- [7] Hiroshi Sawada, Nobutaka Ono, Hirokazu Kameoka, Daichi Kitamura, and Hiroshi Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, pp. e12, 2019.
- [8] Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman, Anees Abrol, Md Abdur Rahaman, Jiayu Chen, et al., "Neuromark: An automated and adaptive ica based pipeline to identify reproducible fmri markers of brain disorders," *NeuroImage: Clinical*, vol. 28, pp. 102375, 2020.
- [9] Shile Qi, Jing Sui, Jiayu Chen, Jingyu Liu, Rongtao Jiang, Rogers Silva, Armin Iraj, Esvar Damaraju, Mustafa Salman, Dongdong Lin, et al., "Parallel group ica+ ica: Joint estimation of linked functional network variability and structural covariation with application to schizophrenia," *Human brain mapping*, vol. 40, no. 13, pp. 3795–3809, 2019.
- [10] Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nature communications*, vol. 12, no. 1, pp. 1–17, 2021.
- [11] Anees Abrol, Manish Bhattarai, Alex Fedorov, Yuhui Du, Sergey Plis, Vince Calhoun, Alzheimer's Disease Neuroimaging Initiative, et al., "Deep residual learning for neuroimaging: an application to predict progression to alzheimer's disease," *Journal of neuroscience methods*, vol. 339, pp. 108701, 2020.
- [12] Ishaan Batta, Anees Abrol, Vince D Calhoun, and Alzheimer's Disease Neuroimaging Initiative, "Svr-based multimodal active subspace analysis for the brain using neuroimaging data," *bioRxiv*, pp. 2022–07, 2022.
- [13] Eric Feczko and Damien A Fair, "Methods and challenges for assessing heterogeneity," *Biological psychiatry*, vol. 88, no. 1, pp. 9–17, 2020.
- [14] Ishaan Batta, Anees Abrol, Zening Fu, Adrian Preda, Theo GM van Erp, and Vince D Calhoun, "Building models of functional interactions among brain domains that encode varying information complexity: A schizophrenia case study," *Neuroinformatics*, vol. 20, no. 3, pp. 777–791, 2022.
- [15] Ishaan Batta, Anees Abrol, Zening Fu, and Vince D Calhoun, "Learning active multimodal subspaces in the brain," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 3822–3825.
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [17] Zhengjia Dai and Yong He, "Disrupted structural and functional brain connectomes in mild cognitive impairment and alzheimer's disease," *Neuroscience Bulletin*, vol. 30, no. 2, pp. 217–232, 2014.