2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385864

Integrating Multimodal Contrastive Learning and Cross-Modal Attention for Alzheimer's Disease Prediction in Brain Imaging Genetics

Rong Zhou[1], Houliang Zhou[1], Li Shen[2], Brian Y. Chen[1], Yu Zhang[3], and Lifang He[1*]
[1]*Department of Computer Science and Engineering, Lehigh University*, Bethlehem, PA, USA
{roz322, hoz421, byc210, yuzi20, lih319}@lehigh.edu
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*, Philadelphia, PA, USA
li.shen@pennmedicine.upenn.edu
[3]*Department of Bioengineering, Lehigh University*, Bethlehem, PA, USA

***Abstract*—High annotation costs serve as a significant hurdle in deploying modern deep learning architectures for clinically relevant medical applications, especially when dealing with the inherent heterogeneity of multimodal data, proving the critical need for innovative algorithms that can effectively utilize unlabeled data. In this paper, we propose a model named MCLCA, which integrates multimodal contrastive learning and cross-modal attention to diagnose Alzheimer's Disease (AD) and identify biomarkers using both labeled and unlabeled multimodal brain imaging genetics data. Through multimodal contrastive learning, MCLCA can effectively learn representations even in the absence of sufficient labels. By utilizing cross-modal attention blocks, the model captures deep connections between different modalities, providing a more comprehensive view of diagnosis. Our proposed MCLCA model is evaluated using the ADNI database with three imaging modalities (VBM-MRI, FDG-PET, and AV45-PET) and genetic SNP data. The results demonstrate that MCLCA can identify important biomarkers with better prediction accuracy compared to the existing methods. The source code is available at https://github.com/MCLCA.**

***Index Terms*—Brain imaging genetics, contrastive learning, cross-modal attention, Alzheimer's disease**

I. Introduction

Alzheimer's disease (AD) is one of the most severe neurodegenerative disorders, profoundly affecting millions of individuals worldwide [1]. In recent years, the convergence of brain imaging genetics has been recognized for its potential in detecting AD, including its early stage, mild cognitive impairment (MCI). Techniques like positron emission tomography (PET) and magnetic resonance imaging (MRI) are increasingly employed to establish connections between brain regions and genetic markers, such as Single Nucleotide Polymorphisms (SNPs). A surge of studies [2]–[4] validates the strong ties between brain imaging traits and genetic factors in AD. This combined approach offers potential for new biomarkers and advances in treatments.

Recent studies have increasingly used deep neural networks for multimodal brain imaging genetics [5]–[12], but they often focus on specific applications or require full supervision, leading to a dependency on phenotypic data like medical history, which may introduce biases. Additionally, challenges in this field include unreliable or absent labels in datasets and data heterogeneity due to different acquisition methods and equipment [13]. These factors make it difficult to compare datasets directly and integrate data effectively. Therefore, there is a critical need to develop robust methods for accurate disease prediction and classification in the diverse landscape of multimodal brain imaging genetics.

Self-supervised contrastive learning emerges as a viable solution to circumvent the above challenges, primarily due to its ability to learn data representations without reliance on labels [14]. Many various methods have been developed and widely recognized. SimCLR [15], for instance, leverages large batch sizes and data augmentation strategies to learn efficient embeddings. BYOL [16] eliminates the necessity of negative samples, introducing an innovative bootstrapped approach. Barlow Twins [17] emphasize reducing redundancy in representations by utilizing a decorrelation loss. SimSiam [18] promotes similarity between two augmented views of the same image without requiring batch normalization. NNCLR [19] extends this idea, using nearest neighbor techniques to further refine the learning. However, while these methods have shown promising results, their primary focus remains on uni-modal data. Such a focus poses limitations when confronted with the inherent heterogeneity of multimodal medical datasets. Recently, ContIG [20] was proposed for multimodal medical imaging with genetics. While this method facilitates multimodal contrastive learning, it sometimes overlooks valuable intra-modality contrastive information, leading to potential losses in representational richness.

In this paper, we propose a model named MCLCA, which integrates multimodal contrastive learning and cross-modal attention to diagnose disease and identify biomarkers using brain imaging genetics data. As depicted in Fig. 1, our model consists of two closely intertwined modules: (i) Multimodal Contrastive Learning (MCL) Module: This module starts by using a deep neural network (DNN) to preliminarily extract features from each modality. Through contrastive learning, the model is then encouraged to learn more comprehensive

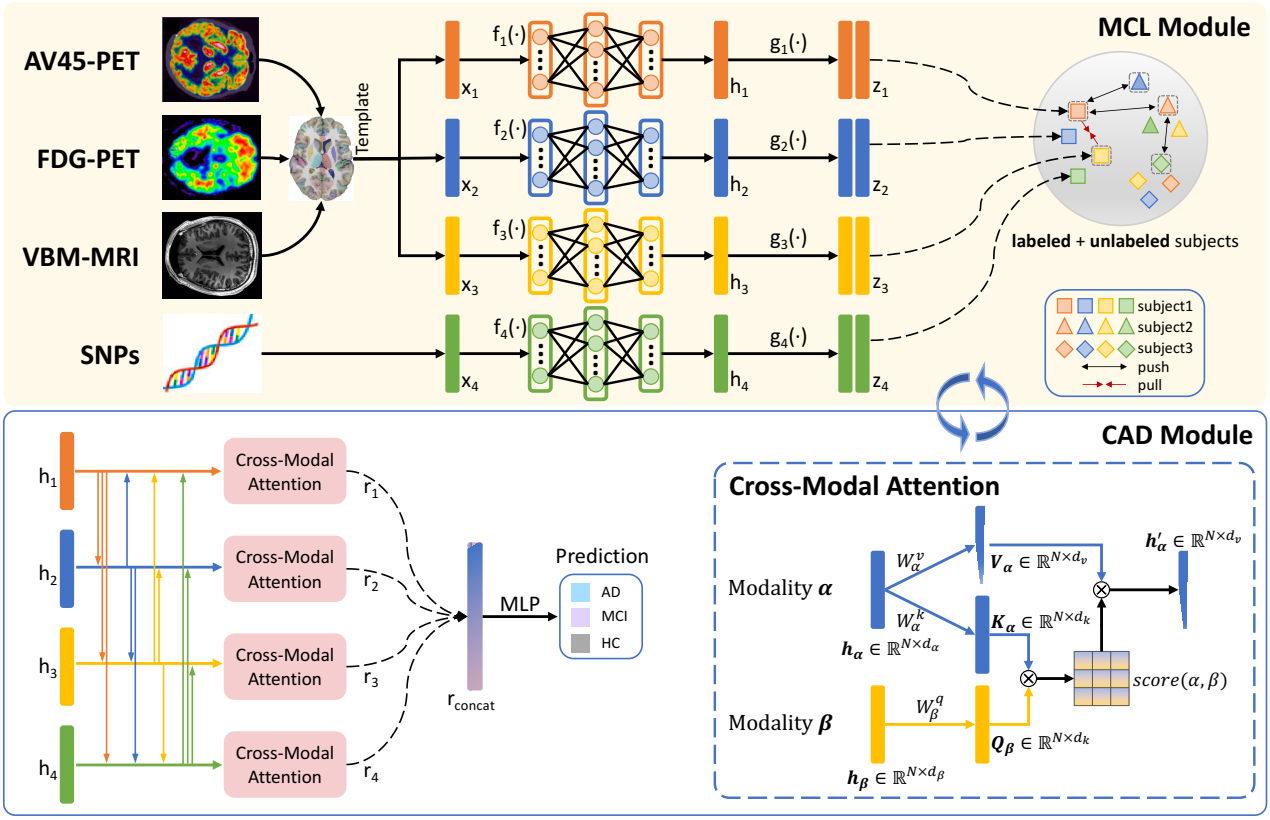

Fig. 1. An overview of the proposed framework. Assuming there are three imaging modalities (AV45-PET, FDG-PET, and VBM-MRI, denoted as x_1, x_2, x_3 respectively) and one genetic modality (SNPs, denoted as x_4). In the multimodal contrastive learning (MCL) module, several deep neural networks $f(\cdot)$ are used to extract the hidden representations h_1, h_2, h_3, h_4. Then the MLP-based projection heads $g(\cdot)$ are employed to produce the representations z_1, z_2, z_3, z_4 for contrastive loss computation. In the cross-modal attention disease prediction (CAD) module, some cross-modal attention blocks are employed for cross-modal information fusion based on the hidden representations h_1, h_2, h_3, h_4. Finally, the cross-modal attention representations r_1, r_2, r_3, r_4 are concatenated to obtain the disease prediction by an MLP.

and discerning hidden representations for each modality. (ii) Cross-Modal Attention Disease Prediction (CAD) Module: We employ cross-modal attention blocks to process the hidden representations acquired from the MCL module. This enables the model to discern deep connections between different modalities, supplying a more holistic view of information for diagnosis.

Our comprehensive experiments, utilizing the real-world ADNI dataset with three imaging modalities (VBM-MRI, FDG-PET, and AV45-PET) and genetic SNP data, demonstrate that our model performs significantly better than other competitive models in classifying AD vs. HC, AD vs. MCI, and MCI vs. HC groups. Furthermore, our model's explanations can highlight disorder-specific biomarkers that align with neuroscience findings. Lastly, we provide evidence that combining classification and correlation models amplifies disease prediction efficacy.

II. Methods

A. Multimodal Contrastive Learning (MCL) Module

Suppose there are N subjects, each with M different modalities. For the i-th subject, the m-th modality is denoted as x_m^i, where $i = 1, 2, \cdots, N$, and $m = 1, 2, \cdots, M$.

Fig. 1 gives an overview of our proposed model that consists of several interrelated phases. The initial step in our process involves the generation of modality-specific hidden representations. For this, we employ the separate deep neural network denoted as $f_m(\cdot)$ for the corresponding modality x_m^i to extract the intrinsic hidden representations expressed as $h_m^i = f_m(x_m^i)$.

After the hidden representations h_m^i are obtained, we further transform them using projection heads, denoted by $g_m(\cdot)$. Each projection head is structured as a nonlinear Multilayer Perceptron (MLP) [21] comprising a single hidden layer. The projection head operates on the hidden representations h_m^i and projects it onto a new representations space, creating the transformed representations denoted as $z_m^i = g_m(h_m^i)$.

One of the core of our approach is the application of contrastive learning [22]. In contrastive learning, the aim is to distinguish between positive and negative pairs. In this paper, we define the positive pairs to be the modality pairs from the same subject. By contrast, negative pairs are defined as the different modality pairs from different subjects.

To guide this learning process, we introduce a loss function

that seeks to minimize the distance between positive pairs while maximizing the distance between negative pairs in the transformed representation space. For clarity, we begin by exploring contrastive loss across two modalities: modality α and modality β. For the i-th subject x_α^i from modality α, we consider the same i-th subject x_β^i from modality β as the positive subject and other subjects x_β^j as the negative subjects. Hence, the contrastive loss $\mathcal{L}_{\text{cont}}(\alpha,\beta)$ comprises two components: (i) $\mathcal{L}(\alpha,\beta)$, which anchors the modality α and contrasts the modality β, and (ii) $\mathcal{L}(\beta,\alpha)$, which anchors the modality β and contrasts the modality α:

$$\begin{aligned}\mathcal{L}(\alpha,\beta) &= -\sum_{i=1}^{N}\log\frac{\exp\left(\cos\left(z_\alpha^i, z_\beta^i\right)/\tau\right)}{\sum_{j=1,j\neq i}^{N}\exp\left(\cos\left(z_\alpha^i, z_\beta^j\right)/\tau\right)}\\ \mathcal{L}_{\text{cont}}(\alpha,\beta) &= \lambda\mathcal{L}(\alpha,\beta)+(1-\lambda)\mathcal{L}(\beta,\alpha),\end{aligned} \tag{1}$$

where τ denotes the temperature coefficient, cos represents the cosine similarity and $\lambda \in [0,1]$ is a hyperparameter used for adjusting the weight of the loss.

We now generalize the contrastive loss from two modalities to a broader multimodal setting. Specifically, we conduct pairwise contrasting across all modalities Hence, the generalized multimodal contrastive loss can be defined as:

$$\mathcal{L}_{\text{mcl}} = \sum_{\alpha=1}^{M-1}\sum_{\beta=\alpha+1}^{M}\mathcal{L}_{\text{cont}}(\alpha,\beta). \tag{2}$$

As the model optimizes this loss function, it is encouraged to generate more similar representations for different modalities of the same subject and distinctly different representations for different subjects, enhancing the model's ability to discriminate effectively among subjects based on multimodal data.

B. Cross-Modal Attention Disease Prediction (CAD) Module

Following the hidden representations h_m obtained from the MCL module, the CAD module takes these hidden representations to further process disease prediction. We employ several cross-modal attention [23] blocks to harmoniously fuse multimodal information, capitalizing on their ability to selectively highlight and integrate crucial inter-modal relationships, thereby maximizing the potential of the combined modalities for more accurate disease prediction.

Suppose there are three linear transformation matrices, $\mathbf{W}_m^q \in \mathbb{R}^{d_m\times d_k}, \mathbf{W}_m^k \in \mathbb{R}^{d_m\times d_k}, \mathbf{W}_m^v \in \mathbb{R}^{d_m\times d_v}$, operating on the m-th modality feature representations $h_m = [h_m^1, h_m^2, \cdots, h_m^N] \in \mathbb{R}^{N\times d_m}$ to obtain: $Q_m = h_m\mathbf{W}_m^q, K_m = h_m\mathbf{W}_m^k, V_m = h_m\mathbf{W}_m^v$. Now we still consider two modalities: modality α and modality β. First, the attention score between modality α and modality β is computed as:

$$\text{score}(\alpha,\beta) = \text{softmax}\left(\frac{Q_\beta K_\alpha^T}{\sqrt{d_k}}\right). \tag{3}$$

Then, the cross-modal attention representations of modality α which considers the information from modality β can be calculated as $h'_\alpha = \text{score}(\alpha,\beta)V_\alpha$.

Considering the information from all other modalities, the cross-modal attention representations of modality α can be computed as:

$$r_\alpha = \sum_{\beta=1,\beta\neq\alpha}^{M} h'_\alpha = \sum_{\beta=1,\beta\neq\alpha}^{M}\text{score}(\alpha,\beta)V_\alpha. \tag{4}$$

We compute cross-modal attention representations for each modality, deriving representations $r_1, r_2, \cdots, r_M$. Then we concatenate these all cross-modal attention representations, formulated as $r_{\text{concat}} = [r_1, r_2, \cdots, r_M]$.

Then r_{concat} is processed through an MLP consisting of a hidden layer with a ReLU activation [24]. The outputs of the MLP then pass through the *softmax* function [25] to derive the disease prediction $\hat{y}$. However, it's important to note that not every subject in our dataset has a disease label. To circumvent this, we introduce a masking strategy.

For the i-th subject, the disease prediction is denoted as $\hat{y}^i$. For each subject, we create a mask denoted as k^i. Mask k^i is either 0 (indicating a missing label) or 1 (indicating the presence of a label). With the ground truth label represented as y^i and disease prediction $\hat{y}$, the masked cross-entropy loss can be defined:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{N} k^i y^i \log\left(\hat{y}^i\right). \tag{5}$$

This loss only considers subjects where the mask k is set to 1 (indicating the presence of a label). Overall, our final training objective can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{mcl}} + \gamma\mathcal{L}_{\text{cls}}, \tag{6}$$

where $\mathcal{L}_{\text{mcl}}$ is the self-supervised multimodal contrastive loss, while $\mathcal{L}_{\text{cls}}$ is the supervised cross-entropy disease prediction loss. The coefficient γ is a tunable hyperparameter, harmonizing the scales of the individual loss components, ensuring an equitable contribution from both during the model optimization process.

III. Experiments and Results

A. Data Acquisition and Preprocessing

Our study utilized brain imaging and genetic data from 887 participants in the ADNI database [26], including 520 with disease labels (120 AD, 251 MCI, and 149 HC) and 367 without specific labels. The data comprised three imaging modalities: structural Magnetic Resonance Imaging (VBM-MRI), 18 F-fluorodeoxyglucose Positron Emission Tomography (FDG-PET), and 18 F-florbetapir PET (AV45-PET). We aligned these images to each participant's visit, standardizing them to the MNI space and segmenting them into 90 Regions of Interest (ROIs) using the AAL-90 atlas [27], focusing on gray matter, white matter, and cerebrospinal fluid maps. For genetic analysis, we selected 54 SNPs near the AD risk gene APOE from the AlzGene database[1], using data genotyped by Illumina platforms and subjected to standard quality control.

[1] www.alzgene.org

TABLE I
CLASSIFICATION PERFORMANCE COMPARISON. THE BEST RESULTS ARE IN BOLD.

Task	Measures	SimCLR	BYOL	SimSam	Barlow Twins	ConIG	(w/o) CL	(w/o) CA	**MCLCA**
AD vs. HC	ACC	0.826±0.047	0.794±0.052	0.847±0.046	0.858±0.044	0.884±0.037	0.883±0.030	0.901±0.027	**0.914±0.021**
	AUC	0.904±0.028	0.896±0.035	0.927±0.022	0.916±0.035	0.934±0.020	0.945±0.023	0.952±0.019	**0.974±0.017**
	Sensitivity	0.862±0.044	0.853±0.047	0.891±0.035	0.857±0.042	0.864±0.038	0.859±0.035	0.874±0.036	**0.898±0.032**
	Specificity	0.741±0.075	0.739±0.084	0.813±0.055	0.746±0.029	0.773±0.057	0.863±0.052	0.886±0.048	**0.918±0.054**
AD vs. MCI	ACC	0.738±0.055	0.717±0.061	0.760±0.061	0.756±0.059	0.771±0.054	0.781±0.039	0.803±0.035	**0.825±0.032**
	AUC	0.812±0.059	0.790±0.066	0.869±0.066	0.865±0.057	0.870±0.057	0.826±0.036	0.845±0.039	**0.889±0.038**
	Sensitivity	0.766±0.061	0.741±0.068	0.815±0.068	0.814±0.049	0.821±0.044	0.797±0.044	0.823±0.031	**0.856±0.029**
	Specificity	0.724±0.083	0.706±0.088	0.723±0.088	0.725±0.065	0.752±0.046	0.709±0.061	0.749±0.058	**0.774±0.047**
HC vs. MCI	ACC	0.594±0.079	0.566±0.081	0.581±0.073	0.601±0.071	0.622±0.068	0.597±0.042	0.627±0.037	**0.712±0.035**
	AUC	0.637±0.045	0.601±0.049	0.626±0.042	0.643±0.046	0.643±0.044	0.692±0.061	0.716±0.058	**0.824±0.049**
	Sensitivity	0.623±0.072	0.603±0.086	0.620±0.075	0.625±0.081	0.633±0.077	0.558±0.105	0.675±0.083	**0.738±0.061**
	Specificity	0.645±0.088	0.621±0.091	0.637±0.135	0.647±0.102	0.617±0.112	0.479±0.104	0.602±0.096	**0.707±0.086**

B. Evaluation of Disease Classification Performance

For our disease prediction evaluation, subjects with disease labels were segmented into three distinct groups: AD vs. HC, AD vs. MCI, and MCI vs. HC. Performance was assessed using four established metrics: accuracy (ACC), area under the curve (AUC), sensitivity, and specificity. Given the limited number of labeled subjects in our study, we implement a 5-fold cross-validation process and present the results as an average alongside the standard deviation. In each fold, all unlabeled subjects were included to ensure a sufficiently large sample size for the self-supervised contrastive learning phase. Subsequently, a masking technique was utilized to allow these unlabeled subjects to participate in the supervised disease prediction. This design aimed to maximize the utilization of available data while maintaining methodological rigor. To benchmark the efficacy of our approach, we utilized state-of-the-art contrastive methods as baselines, including SimCLR [15], BYOL [16], Barlow Twins [17], SimSiam [18], NNCLR [19], and ContIG [20]. While our method integrates cross-modal attention followed by an MLP for disease prediction, these baseline models utilized a DNN classifier for the same purpose. To ensure a fair comparison, all methods were evaluated under the same experimental conditions, using identical data splits and configurations. During the experiments, we employed the Adam optimizer for model training. More specifically, we configured the learning rate to be 0.0001 and set the weight decay parameter to 0.001. Additionally, the value for the parameter γ in our model was empirically set to 2, and we trained the model for 1000 epochs.

Table I presents the classification performance across different tasks, where $\pm$ represents the standard deviation of evaluation scores across the 5 folds. Notably, our proposed MCLCA method consistently outperforms other state-of-the-art contrastive methods in all metrics. In particular, for the AD vs. HC, AD vs. MCI, and HC vs. MCI tasks, MCLCA exhibits clear superiority. While methods such as SimSam and Barlow Twins provide competitive results, especially in AD vs. HC and AD vs. MCI tasks, MCLCA still achieves the highest scores. The margin of improvement demonstrates the value of our novel approach, especially in the challenging HC vs. MCI task where traditional methods show weaker differentiation capabilities. A discernible trend is a decrease in ACC and AUC values from the AD vs. HC task to the HC vs. MCI task for all methods, reflecting the challenge of distinguishing between closely related disease states. Overall, the results highlight the efficacy and robustness of our MCLCA method in disease classification using multimodal data.

Our proposed MCLCA model contains two important components: Contrastive Learning (CL) and Cross-Modal Attention (CA). To understand the impact of each component, we conducted ablation studies by evaluating the two additional models: the MCLCA model trained without CL (w/o CL) and without the CA (w/o CA). Table I presents the results for the AD vs. HC task, it is evident that the removal of either component results in a performance drop: from a peak accuracy of 91.4% in the full model to 88.3% without CL and 90.1% without CA. Similar trends are observed in AD vs. MCI and MCI vs. HC tasks, with CL playing a crucial role in distinguishing related classes and CA enhancing feature interaction across modalities. The results confirm the combined importance of CL and CA in achieving state-of-the-art classification performance.

C. The Most Discriminative Brain Regions and SNPs

Determining the most significant brain regions, and Single Nucleotide Polymorphisms (SNPs) stands as a pivotal task in the accurate diagnosis of AD. To achieve this, we utilized the integrated gradients interface from Captum [28]. This interface offers the capability to allocate importance scores to each feature from various modalities. It analyzes the pre-trained model, shedding light on the intricate relationship between input features and their influence on the final predictions.

Figs. 2(a-c) shows the top 20 discriminative ROIs identified by the proposed method from each individual brain imaging modality. Fig. 2(d) shows the top 20 discriminative ROIs selected by the three modalities together. Here, specific regions like the hippocampus, parahippocampal gyrus, precuneus, and temporal lobe regions are underscored, confirming their significance. Past research has highlighted the crucial role these regions play in AD and MCI [29]–[31]. Moreover, the

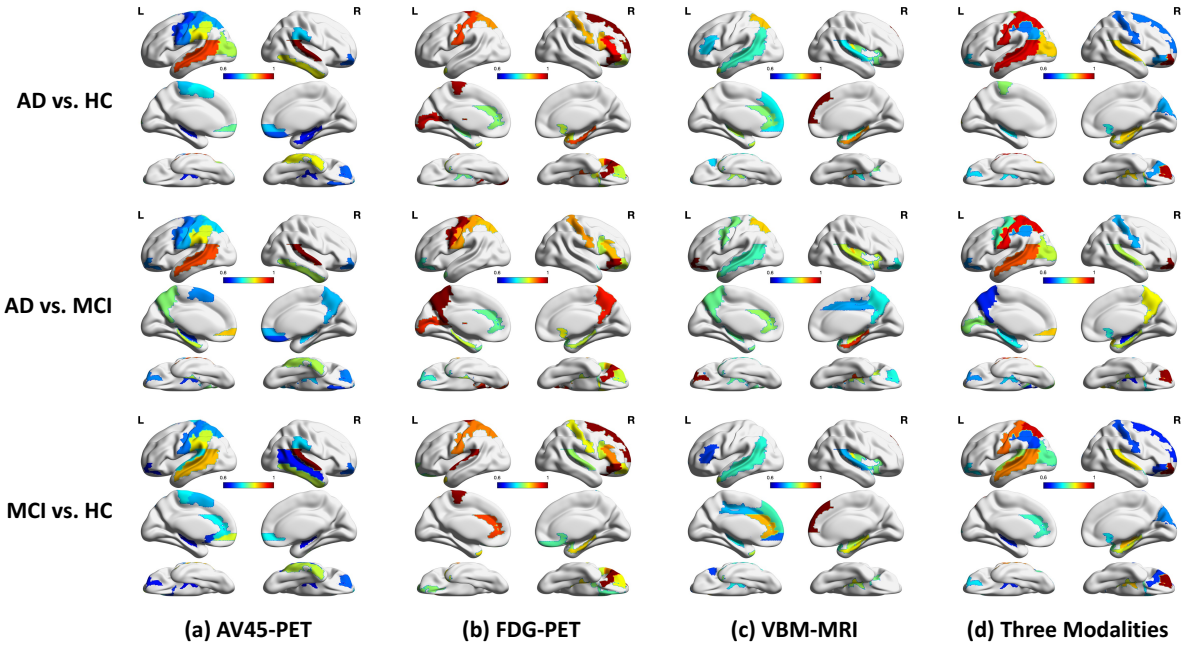

Fig. 2. Top 20 discriminative ROIs identified by MCLCA from three brain imaging modalities for three different classification groups in lateral, medial, and ventral view. The color bar indicates the importance score.

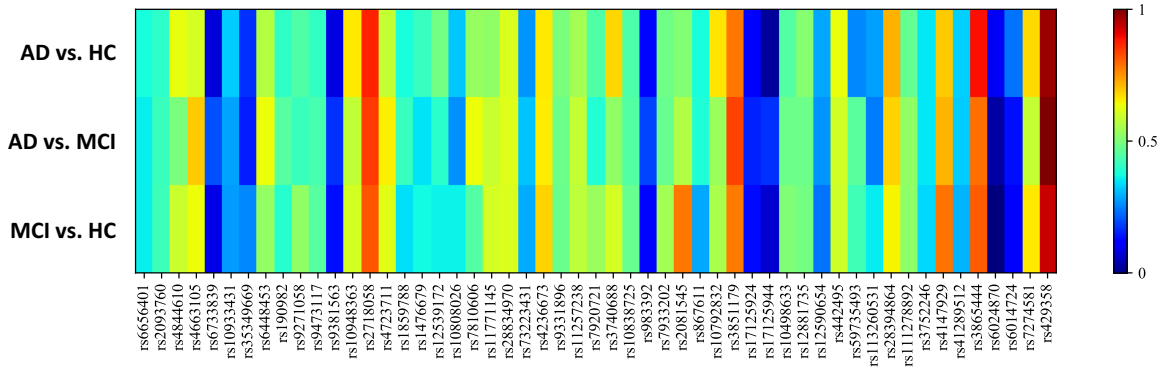

Fig. 3. The importance scores of SNPs. The red color indicates a high score.

selection of ROIs exhibits variation across different classification groups, underscoring the adaptability of our model in pinpointing the critical ROIs pertinent to distinct diseases.

Moving to genetic markers, Fig. 3 highlights key SNPs like rs429358, rs2718058, rs3851179, and rs3865444, identified for their high discriminatory power. Their significance aligns with previous genetic research, underlining the precision and relevance of our approach [32], [33].

IV. Conclusion

In this paper, we introduce MCLCA, a multimodal model that leverages the capabilities of multimodal contrastive learning and cross-modal attention on a combination of labeled and unlabeled data. We showcase its potential in both diagnosing AD and identifying crucial biomarkers. Our experiments using the ADNI dataset encompassing three imaging modalities and genetic SNP data confirmed the superiority of MCLCA over other state-of-the-art contrastive methods. Specifically, MCLCA not only demonstrates consistent superiority over these techniques in AD disease classification tasks but also excels in distinguishing between similar disease states like HC and MCI. Moreover, MCLCA is able to identify the significant brain regions related to AD, such as the hippocampus and temporal lobes, alongside significant SNPs like rs429358, which aligns with existing literature. The alignment between our findings and those of previous studies bolsters the credibility of MCLCA, demonstrating its potential as a valuable tool in the realm of AD research. Future work could explore additional imaging techniques for deeper AD insights, assess the model's scalability and applicability to diverse populations, and integrate temporal analysis to study AD's progression.

ACKNOWLEDGMENT

This work is partially supported by the National Institutes of Health (U01AG068057, U01AG-066833, R01LM013463, R01MH129694, R21MH130956, R21AG080425 and R21EY034179), National Science Foundation (MRI-2215789, IIS-1909879 and IIS-2319451), Alzheimer's Association grant (AARG-22-972541), and Lehigh's grants under Accelerator (S00010293), CORE (001250), and FIG (FIGAWD35).

REFERENCES

[1] M. Catania, L. Colombo, S. Sorrentino, A. Cagnotto, J. Lucchetti, M. C. Barbagallo, I. Vannetiello, E. R. Vecchi, M. Favagrossa, M. Costanza *et al.*, "A novel bio-inspired strategy to prevent amyloidogenesis and synaptic damage in alzheimer's disease," *Molecular psychiatry*, pp. 1–8, 2022.

[2] L. Shen and P. M. Thompson, "Brain imaging genetics: integrated analysis and machine learning," in *IEEE International conference on bioinformatics and biomedicine*. IEEE Computer Society, 2021, pp. 1–1.

[3] Y. Xin, J. Sheng, M. Miao, L. Wang, Z. Yang, and H. Huang, "A review of imaging genetics in alzheimer's disease," *Journal of clinical neuroscience*, vol. 100, pp. 155–163, 2022.

[4] H. Zhou, L. He, Y. Zhang, L. Shen, and B. Chen, "Interpretable graph convolutional network of multi-modality brain imaging for alzheimer's disease diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[5] L. Du, K. Liu, X. Yao, S. L. Risacher, J. Han, A. J. Saykin, L. Guo, and L. Shen, "Detecting genetic associations with brain imaging phenotypes in alzheimer's disease via a novel structured scca approach," *Medical image analysis*, vol. 61, p. 101656, 2020.

[6] X. Wang, R. Zhou, K. Zhao, A. Leow, Y. Zhang, and L. He, "Normative modeling via conditional variational autoencoder and adversarial learning to identify brain dysfunction in alzheimer's disease," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–4.

[7] S. Ghosal, Q. Chen, G. Pergola, A. L. Goldman, W. Ulrich, D. R. Weinberger, and A. Venkataraman, "A biologically interpretable graph convolutional network to link genetic risk pathways and imaging phenotypes of disease," in *International conference on learning representations*, 2022.

[8] H. Zhou, Y. Zhang, B. Y. Chen, L. Shen, and L. He, "Sparse interpretation of graph convolutional networks for multi-modal diagnosis of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 469–478.

[9] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific reports*, vol. 11, no. 1, p. 3254, 2021.

[10] M.-L. Wang, W. Shao, X.-K. Hao, and D.-Q. Zhang, "Machine learning for brain imaging genomics methods: A review," *Machine intelligence research*, vol. 20, no. 1, pp. 57–78, 2023.

[11] R. Zhou, H. Zhou, B. Y. Chen, L. Shen, Y. Zhang, and L. He, "Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 681–691.

[12] K. Zhang, J. Yu, Z. Yan, Y. Liu, E. Adhikarla, S. Fu, X. Chen, C. Chen, Y. Zhou, X. Li *et al.*, "Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks," *arXiv preprint arXiv:2305.17100*, 2023.

[13] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of the royal society interface*, vol. 15, no. 141, p. 20170387, 2018.

[14] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. Proceedings of machine learning research, 2020, pp. 1597–1607.

[16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. Proceedings of machine learning research, 2021, pp. 12 310–12 320.

[18] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 15 750–15 758.

[19] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9588–9597.

[20] A. Taleb, M. Kirchler, R. Monti, and C. Lippert, "Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 908–20 921.

[21] A. Zell, *Simulation neuronaler netze*. Addison-Wesley Bonn, 1994, vol. 1, no. 5.3.

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[23] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning*, 2010, pp. 807–814.

[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[26] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The alzheimer's disease neuroimaging initiative," *Neuroimaging clinics*, vol. 15, no. 4, pp. 869–877, 2005.

[27] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[28] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[29] Y. Mu and F. H. Gage, "Adult hippocampal neurogenesis and its role in alzheimer's disease," *Molecular neurodegeneration*, vol. 6, no. 1, pp. 1–9, 2011.

[30] G. W. Van Hoesen, J. C. Augustinack, J. Dierking, S. J. Redman, and R. Thangavel, "The parahippocampal gyrus in alzheimer's disease: clinical and preclinical neuroanatomical correlates," *Annals of the New York academy of sciences*, vol. 911, no. 1, pp. 254–274, 2000.

[31] G. Karas, P. Scheltens, S. Rombouts, R. Van Schijndel, M. Klein, B. Jones, W. Van Der Flier, H. Vrenken, and F. Barkhof, "Precuneus atrophy in early-onset alzheimer's disease: a morphometric structural mri study," *Neuroradiology*, vol. 49, pp. 967–976, 2007.

[32] I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu *et al.*, "Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk," *Nature genetics*, vol. 51, no. 3, pp. 404–413, 2019.

[33] J. Zhen, X. Huang, N. Van Halm-Lutterodt, S. Dong, W. Ma, R. Xiao, and L. Yuan, "Apoe rs429358 and rs7412 polymorphism and gender differences of serum lipid profile and cognition in aging chinese population," *Frontiers in aging neuroscience*, vol. 9, p. 248, 2017.