Multimodal Contrastive Learning for Alzheimer's Disease Prediction in Imaging Genetics

Jonas Fallmann
Institute for Machine Learning
Johannes Kepler University
Linz
jonas.fallmann@jku.at

Erich Kobler
Institute for Machine Learning,
LIT AI Lab,
Department of Virutal Morphology,
Clinical Research Institute Medical AI
Johannes Kepler University
Linz
erich.kobler@jku.at

Abstract

Alzheimer's disease (AD) is an inherently multimodal pathology driven by complex genetic and phenotypic interactions, making reliable early detection a critical challenge. Existing multimodal approaches often struggle to effectively align static baseline genetic risk with longitudinal physical changes. In this work, we introduce a novel two-stage contrastive learning framework integrating Single Nucleotide Polymorphisms (SNPs) and structural MRI volumes. To overcome the bottleneck of single-timepoint genetic measurements, we propose an age-conditioned augmentation strategy that generates time-aware genetic embeddings for longitudinal contrastive pairing. Utilizing a dynamic Gated Fusion mechanism for downstream classification, our approach effectively weights modality contributions. Evaluated on the ADNI database, our framework consistently outperforms strong classical baselines and state-of-the-art generative models, demonstrating particularly significant improvements in early-stage cognitive decline detection.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the leading global cause of dementia [11]. Because current treatments primarily slow progression rather than cure it [16], early detection is critical. Capturing the complete disease trajectory requires integrating upstream genetic susceptibility with downstream phenotypic consequences.

Existing multimodal AD research predominantly fuses imaging modalities such as MRI and PET using standard supervised architectures [1, 13, 14, 10]. However, these methods often struggle to learn robust representations when labeled data is scarce. While contrastive learning has emerged as a powerful alternative for extracting features from medical imaging [8, 9, 5] and imaging-genetics cohorts [15, 12, 17], its potential to align high-dimensional SNPs with MRI brain volumes for AD remains underexplored. A primary hurdle is data asymmetry: effectively combining a single static genetic measurement with longitudinal imaging data requires complex setups that standard fusion approaches fail to address [6].

In this work, we introduce a multimodal contrastive learning architecture to extract meaningful, modality-invariant latent representations from MRI brain volumes and SNPs. To overcome the bottleneck of relying on a single genetic measurement per subject, we introduce time-aware genetic embeddings. By augmenting static genetic profiles with the patient's age at the time of the scan, we enable the generation of rich, multi-view positive pairs. We then evaluate the expressivity of

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

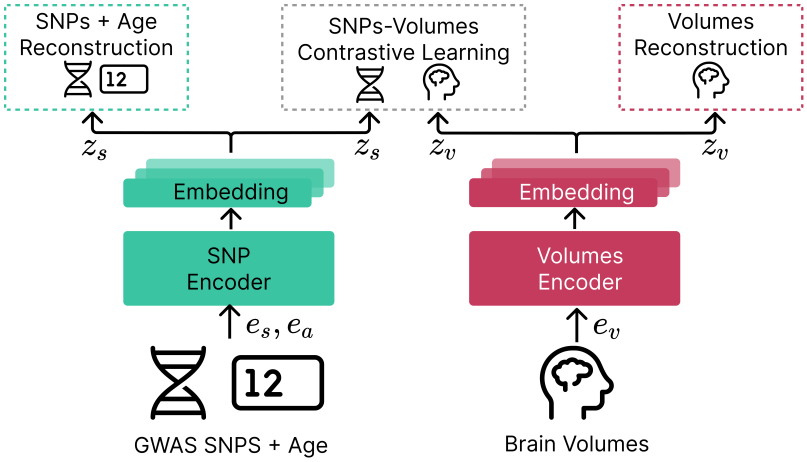

Figure 1: Contrastive learning setup featuring dual-branch encoders for input modalities and distinct decoders for feature-preserving reconstruction. Inter-modal contrastive loss aligns patient-specific embeddings across modalities.

these learned embeddings by training a downstream classifier on the frozen encoders to predict AD diagnosis and cognitive decline.

2 Methodology

To address the problem of Alzheimer's disease classification, we adopt a two-stage pipeline that decouples representation learning from downstream classification. In the first stage, a contrastive encoder module is trained to produce dense latent representations that capture the most salient and modality-invariant features of the input data. In the second stage, a fusion model combines the latent representation produced by the frozen contrastive encoding module. Classification of the disease class and regression of the Mini-Mental State Examination (MMSE) [4] score are performed.

2.1 Contrastive encoding

At a conceptual level, the contrastive encoding module maps different data modalities into a shared latent space (Figure 2). Using contrastive learning, representations of the same subject at a specific time point are pulled together, while representations of different entities are pushed apart, promoting modality-invariant embeddings.

Following the SimCLR framework [2], we define the contrastive objective using the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss. Let e_s denote the SNP features, e_a the patient age, and e_v the standardized MRI volumes. We define the modality-specific latent representations as $z_s = f_\theta(e_s, e_a)$ and $z_v = g_\phi(e_v)$, where f_θ and g_ϕ are the genetic and volume encoders, respectively. For a batch of size N, the loss for a genetic anchor $z_s^{(i)}$ against the volume representations $\mathcal{Z}_v$ is formulated as:

$$\ell(z_s^{(i)}, \mathcal{Z}_v) = -\log \frac{\exp(\text{sim}(z_s^{(i)}, z_v^{(i)})/\tau)}{\sum_{k\neq i}^{N} \exp(\text{sim}(z_s^{(i)}, z_v^{(k)})/\tau)}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity and τ represents the temperature. To ensure bidirectional alignment, we symmetrically compute the loss for the volume anchor against the genetic representations, yielding the total contrastive objective $\mathcal{L}_{\text{cont}} = \frac{1}{2N}\sum_{i=1}^{N}[\ell(z_s^{(i)}, \mathcal{Z}_v) + \ell(z_v^{(i)}, \mathcal{Z}_s)]$. In addition, we concurrently apply a mean squared error (MSE) reconstruction objective to preserve features critical for downstream classification thereby enhancing the overall representational quality. The total loss is a weighted sum of the individual loss terms.

To form contrastive pairs, we match a subject's static genetic profile with brain volumes from their longitudinal MRI scans. Because genetic profiles are static and measured only once, we face a critical bottleneck in generating the abundant, distinct multi-view positive pairs required for

effective contrastive learning. To overcome this, we augment the static genetic profile by directly concatenating the subject's scalar age at the time of MRI acquisition. Empirical evaluations showed no significant performance difference between simple concatenation and mapping age through a learned embedding, so we opted for the computationally simpler strategy. This simple augmentation extracts age-conditioned representations, transforming a single genetic profile into multiple distinct views corresponding to each longitudinal scan. Biologically, this strategy is highly plausible, as the phenotypic effects of genetic variants on brain morphology are strongly age-dependent.

2.2 Gated fusion classification

To fuse the projected genetic (z_s) and structural (z_v) latent vectors, we employ a learned modality-level gating mechanism. Scalar modality-specific attention weights are computed via a linear projection and softmax activation:

$$[\alpha_s, \alpha_v] = \text{Softmax}(W_g[z_s, z_v] + b_g)$$

where $[\cdot, \cdot]$ denotes concatenation and $\alpha_s, \alpha_v \in \mathbb{R}$. The final representation is the weighted sum $z_{\text{fused}} = \alpha_s z_s + \alpha_v z_v$, enabling adaptive downweighting of noisy or missing inputs. Let e_d represent the demographic embedding encompassing sex and education years. The fused vector is concatenated to form the complete patient representation $h = [z_{\text{fused}}, e_d]$. This representation is then routed to parallel MLP heads to yield the final predictions $\hat{y} = \text{MLP}(h)$ for diagnostic classification and continuous MMSE score regression, optimized simultaneously via a weighted compound loss.

3 Experimental setup

Our dataset comprised 844 subjects from the ADNI database, yielding 3,784 multimodal pairs. Data curation involved dropping missing records and applying a 90-day soft-alignment threshold to match imaging assessments with the closest temporal diagnostic labels. Across all longitudinal visits, the distribution was 1,354 Cognitively Normal (CN), 1,921 Mild Cognitive Impairment (MCI), and 509 Alzheimer's Disease (AD) assessments. The final-visit classification subset comprised 299 CN, 349 MCI, and 196 AD subjects. We evaluated the framework across three binary classification tasks, a multiclass task (CN vs. MCI vs. AD), and continuous regression of the Mini-Mental State Examination (MMSE) score.The MMSE is a 30-point scale where a score of 30 indicates no cognitive impairment, which we min-max normalized to a [0, 1] range for our prediction task. For the imaging modality, structural MRI volumes were extracted via Freesurfer [3] and directly z-score standardized to preserve absolute volume information, rather than normalizing by intracranial volume. For the genetic modality, SNPs derived from microarray data were ranked by p-value using the Alzheimer's Disease Variant Portal (ADVP) [7]—an aggregated database of multiple studies. These selected SNPs were mapped from discrete categorical values into continuous representation embeddings using the population-based probability transformation matrix introduced by Ko et al. [6]. Both continuous modality-specific embeddings were subsequently fed into our contrastive encoding architecture.

To prevent longitudinal data leakage, we enforced a strict subject-wise split that remained identical across both training stages. Within each cross-validation fold, the contrastive encoder was pretrained using all available timepoints of the training subjects only. Subsequently, the downstream Gated Fusion classifier was trained and evaluated exclusively on the final timepoint of the respective train and test subjects. Performance was assessed using five-fold nested stratified cross-validation. We benchmarked our framework against classical machine learning algorithms trained on concatenated features—among which Random Forest proved to be the strongest performing—as well as a state-of-the-art multimodal generative-adversarial attention framework [6]. Hyperparameters were rigorously optimized within the inner cross-validation loop using exhaustive grid search for classical models and Optuna Bayesian optimization for the neural architectures.

4 Results

Our complete framework, integrating contrastive pretraining with downstream Gated Fusion, demonstrated strong performance across diagnostic classification tasks, particularly when compared to robust classical baselines (Table 1). While we included a recent generative VAE+GAN model as a

baseline, its adversarial training proved highly unstable in practice, frequently underperforming a standard Random Forest Classifier (RFC). Consequently, we treat the classical RFC as our primary competitive benchmark.

Our method achieved the most significant gains in the challenging task of early cognitive decline recognition (CN vs. MCI). The framework reached an Area Under the Curve (AUC) of 0.707, providing a definitive improvement over both the classical RFC (AUC 0.665) and the VAE+GAN (AUC 0.624). However, this classification advantage comes with a clear trade-off in continuous Mini-Mental State Examination (MMSE) regression. The Random Forest Regressor (RFR) achieved lower Root Mean Square Errors (RMSE) than our model. This highlights a limitation of our multi-task compound loss: while joint optimization benefits categorical disease staging, it slightly compromises the precision of continuous cognitive score prediction compared to a dedicated, single-objective regressor.

Ablation experiments confirmed that structural MRI volumes primarily drive predictions in advanced disease stages (AUC of 0.928 for CN vs. AD using MRIa alone), whereas genetic features (SNPs) supply complementary variance for early-stage detection. Crucially, relying solely on either the reconstruction or contrastive objectives yielded suboptimal performance (AUCs of 0.916 and 0.882 for CN vs. AD, respectively). This demonstrates that neither objective alone fully captures multi-modal complexity; rather, their synergy is essential for aligning distinct views in the latent space, noticeably boosting classification across all categories.

Table 1: Comparison of model performance across diagnostic groups. Best results are in bold (highest ROC-AUC, lowest RMSE) and second-best results are underlined.

Model	**Metric**	**CN/AD**	**CN/MCI**	**MCI/AD**	**CN/MCI/AD**
Only Reconstruction	AUC	0.916 ± 0.024	0.668 ± 0.032	0.817 ± 0.035	0.760 ± 0.019
	RMSE	0.133 ± 0.015	$\underline{0.073} \pm 0.006$	0.142 ± 0.010	0.134 ± 0.016
Only Contrastive	AUC	0.882 ± 0.027	0.659 ± 0.046	0.805 ± 0.025	0.739 ± 0.024
	RMSE	0.132 ± 0.009	0.074 ± 0.011	0.140 ± 0.013	0.125 ± 0.009
Single modal SNPs	AUC	0.638 ± 0.069	0.626 ± 0.038	0.534 ± 0.045	0.590 ± 0.028
	RMSE	0.174 ± 0.018	0.103 ± 0.039	0.167 ± 0.024	0.148 ± 0.018
Single modal MRI	AUC	$\underline{0.928} \pm 0.028$	$\underline{0.699} \pm 0.024$	$\underline{0.819} \pm 0.029$	$\underline{0.774} \pm 0.013$
	RMSE	0.128 ± 0.007	0.156 ± 0.011	$\underline{0.136} \pm 0.006$	0.132 ± 0.011
RFC / RFR	AUC	0.893 ± 0.030	0.665 ± 0.037	0.812 ± 0.041	0.747 ± 0.028
	RMSE	$\mathbf{0.125} \pm 0.010$	$\mathbf{0.062} \pm 0.006$	0.137 ± 0.006	$\mathbf{0.114} \pm 0.006$
VAE + GAN Fusion [6]	AUC	0.861 ± 0.023	0.624 ± 0.042	0.768 ± 0.027	0.691 ± 0.022
	RMSE	0.145 ± 0.013	0.077 ± 0.005	$\mathbf{0.133} \pm 0.011$	$\underline{0.123} \pm 0.010$
Contrastive + Fusion (ours)	AUC	$\mathbf{0.936} \pm 0.017$	$\mathbf{0.707} \pm 0.035$	$\mathbf{0.831} \pm 0.019$	$\mathbf{0.778} \pm 0.014$
	RMSE	$\underline{0.127} \pm 0.015$	0.080 ± 0.014	0.144 ± 0.012	0.127 ± 0.006

5 Conclusion

We presented a two-stage multimodal contrastive learning framework for AD detection. By integrating SNPs with longitudinal MRI via an age-conditioned augmentation strategy, our approach mitigates data asymmetry and captures complex disease trajectories. Dynamic Gated Fusion adaptively weights these modalities, outperforming classical baselines particularly in early-stage detection (CN vs. MCI). Ablations confirm that combining contrastive and reconstruction objectives preserves critical pathological details. Future work will integrate PET imaging and fluid biomarkers, validating the framework on independent cohorts.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from the DFG within the SPP2298 under project number 543939932 and from the Austrian Science Fund (FWF) project number 10.55776/COE12.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Giovanna Castellano, Andrea Esposito, Eufemia Lella, Graziano Montanaro, and Gennaro Vessio. Automated detection of Alzheimer's disease: A multi-modal approach with 3D MRI and amyloid PET. *Scientific Reports*, 14(1):5210, 2024. doi: 10.1038/s41598-024-56001-9.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 1597–1607. PMLR, 2020.

[3] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

[4] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975. doi: 10.1016/0022-3956(75)90026-6.

[5] Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935, 2023.

[6] Wonjun Ko, Wonsik Jung, Eunjin Jeon, and Heung-Il Suk. A deep generative-discriminative learning for multimodal representation in imaging genetics. *IEEE Transactions on Medical Imaging*, 41(9):2348–2359, 2022. doi: 10.1109/TMI.2022.3162870.

[7] Pavel P Kuksa, Chia-Lun Liu, Wei Fu, Liming Qu, Yi Zhao, Zivadin Katanic, Kaylyn Clark, Amanda B Kuzma, Pei-Chuan Ho, Kai-Teh Tzeng, et al. Alzheimer's disease variant portal: A catalog of genetic findings for Alzheimer's disease. *Journal of Alzheimer's Disease*, 86(1):461–477, 2022. doi: 10.3233/JAD-215055.

[8] Min Gu Kwak, Yi Su, Kewei Chen, David Weidman, Teresa Wu, Fleming Lure, Jing Li, and Alzheimer's Disease Neuroimaging Initiative. Self-supervised contrastive learning to predict the progression of Alzheimer's disease with 3D amyloid-PET. *Bioengineering*, 10(10):1141, 2023. doi: 10.3390/bioengineering10101141.

[9] Jiaguang Li, Ying Wei, Chuyuan Wang, Qian Hu, Yue Liu, and Long Xu. 3D CNN-based multi-channel contrastive learning for Alzheimer's disease automatic diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022. doi: 10.1109/TIM.2022.3162265.

[10] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius, and Sanjay Misra. Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images. *Journal of Medical and Biological Engineering*, 43(3):291–302, 2023. doi: 10.1007/s40846-023-00801-3.

[11] Saeid Safiri, Amir Ghaffari Jolfayi, Asra Fazlollahi, Soroush Morsali, Aila Sarkesh, Amin Daei Sorkhabi, Behnam Golabi, Reza Aletaha, Kimia Motlagh Asghari, Sana Hamidi, et al. Alzheimer's disease: A comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. *Frontiers in Medicine*, 11:1474043, 2024. doi: 10.3389/fmed.2024.1474043.

[12] Daniel Sens, Liubov Shilova, Adrian V Dalca, Julia Schnabel, and Francesco Paolo Casale. GEMCONT: Genetics-based multimodal contrastive learning for disease-focused imaging genetics. In *Medical Imaging with Deep Learning*, 2026. URL https://openreview.net/forum?id=Y8gkT7s44N.

[13] Juan Song, Jian Zheng, Ping Li, Xiaoyuan Lu, Guangming Zhu, and Peiyi Shen. An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Frontiers in Digital Health*, 3, 2021. doi: 10.3389/fdgth.2021.637386.

[14] Jinju Sun, Chao Cong, Xinpeng Li, Weicheng Zhou, Renxiang Xia, Huan Liu, Yi Wang, Zhiqiang Xu, and Xiao Chen. Identification of Parkinson's disease and multiple system atrophy using multimodal PET/MRI radiomics. *European Radiology*, 34(1):662–672, 2024. doi: 10.1007/s00330-023-10003-9.

[15] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph 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*, pages 20908–20921, 2022.

[16] Sebastian Walsh, Richard Merrick, Edo Richard, Shirley Nurock, and Carol Brayne. Lecanemab for Alzheimer's disease. *BMJ*, 379:o3010, 2022. doi: 10.1136/bmj.o3010.

[17] Rong Zhou, Houliang Zhou, Li Shen, Brian Y. Chen, Yu Zhang, and Lifang He. Integrating multimodal contrastive learning and cross-modal attention for Alzheimer's disease prediction in brain imaging genetics. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pages 1806–1811, 2023. doi: 10.1109/BIBM58861.2023.10385864.