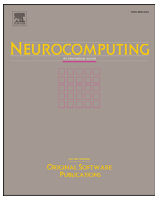
Neurocomputing 694 (2026) 133916

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Foundation models meet multimodal neuroimaging: A generative transformer-based framework for Alzheimer's disease diagnosis

Luca Zedda*, Andrea Loddo, Cecilia Di Ruberto, Alzheimer's Disease Neuroimaging Initiative[1]

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Communicated by Q. Huang

Keywords:
Multimodal learning
Foundation models
Intelligent decision support
Neurodegenerative diseases
Alzheimer's disease
Diffusion models
ADNI

ABSTRACT

Incomplete neuroimaging data remains a major challenge in Alzheimer's disease diagnosis, as many patients undergo only a subset of recommended imaging protocols. This work addresses this limitation by proposing a generative transformer-based framework designed to support multimodal analysis in the presence of missing modalities. We systematically investigate multimodal performance and fairness within a unified foundation model framework for Alzheimer's disease classification while introducing a generative approach that combines structural MRI, DTI, and PET data and leverages ControlNet-based diffusion models to synthesize anatomically consistent surrogate modalities when data are unavailable. These synthetic images are used exclusively as a training-time augmentation strategy for incomplete-modality settings, rather than as replacements for clinical acquisitions. Vision transformers adapted via Low-Rank Adaptation are employed for efficient feature extraction, while clinical variables are integrated through a dedicated projection module. Experimental results show that a transformer-based fusion head can improve upon simple aggregation strategies in some complex multimodal settings, achieving an F1-score of 57.8% in multiclass classification when combined with generative augmentation and clinical data. However, these benefits are not uniform since strong unimodal volumetric PET baselines remain superior in the best-case binary setting, and the effect of generative augmentation is strongly configuration-dependent, with some settings benefiting while others degrading substantially under non-selective synthetic augmentation.

1. Introduction

Neurodegenerative diseases represent a major challenge in geriatrics, both from a clinical and societal perspective. With the growing prevalence of conditions such as Alzheimer's disease (AD) and other forms of dementia, there is an urgent need for robust diagnostic and prognostic tools that can leverage the wealth of multimodal data now available [1,2]. The Alzheimer's Disease Neuroimaging Initiative (ADNI) has emerged as a cornerstone dataset in this domain, offering structural and functional neuroimaging alongside diffusion imaging, positron emission tomography (PET), and comprehensive clinical and demographic assessments [3,4]. Despite its widespread adoption, several methodological challenges remain unresolved, particularly in relation to fairness, modality integration, and handling incomplete data [5].

This article explores three critical aspects of the analysis of neurodegenerative diseases in the context of geriatrics. First, we address the issue of *fairness* in classification tasks based on ADNI, recognizing that demographic and acquisition-related biases may inadvertently influence predictive performance and hinder the generalizability of models across patient subgroups [5,6]. Second, we investigate the relative contributions and synergies of different modalities, including structural magnetic resonance imaging (MRI), PET, diffusion tensor imaging (DTI), and clinical variables by comparing unimodal and multimodal learning strategies. Such analyses provide insight into how different sources

* Corresponding author.
Email addresses: luca.zedda@unica.it (L. Zedda), andrea.loddo@unica.it (A. Loddo), cecilia.dir@unica.it (C. Di Ruberto).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neucom.2026.133916
Received 10 December 2025; Received in revised form 24 April 2026; Accepted 8 May 2026
Available online 9 May 2026

Acronyms

DL Deep Learning
AD Alzheimer's Disease
ADNI Alzheimer's Disease Neuroimaging Initiative
MRI Magnetic Resonance Imaging
PET Positron Emission Tomography
DTI Diffusion Tensor Imaging
MCI Mild Cognitive Impairment
CN Cognitively Normal (controls)
MC Multiclass (classification)
B Binary (classification)
LoRA Low-Rank Adaptation
ControlNet Conditional diffusion model for image-to-image translation
VAE Variational Autoencoder
GAN Generative Adversarial Network
SSIM Structural Similarity Index Measure
PSNR Peak Signal-to-Noise Ratio
MSE Mean Squared Error
RMSE Root Mean Squared Error
MAE Mean Absolute Error
NCC Normalized Cross-Correlation
ADNI1 ADNI Phase 1
ADNI2 ADNI Phase 2
ADNI-GO ADNI-GO Phase
ADNI3 ADNI Phase 3
ADNI4 ADNI Phase 4
PTDEMOG ADNI Demographics Form
AMAS General Knowledge and Cognitive Assessment Questionnaire
NRRP National Recovery and Resilience Plan
MUR Italian Ministry of University and Research
EU European Union
CLS Classification Token

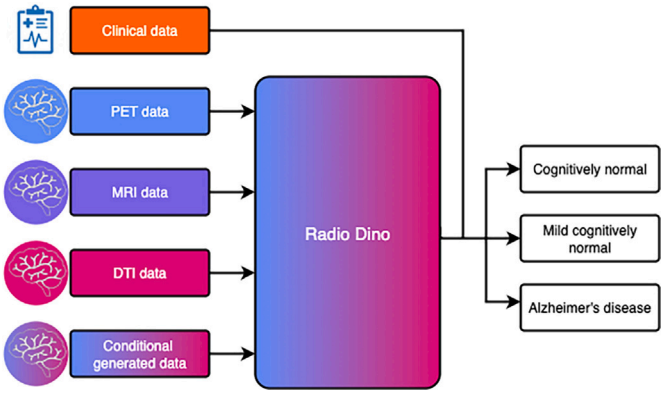


Fig. 1. A schematic of the proposed multimodal analysis pipeline. The framework integrates foundation models with a generative diffusion model for data augmentation and the possibility of clinical data fusion for a complete patient analysis.

of information complement one another in the detection and monitoring of neurodegenerative processes. Finally, we propose the use of image translation diffusion-based models to address the challenge of missing data, which often plays a critical role [7,8]. By learning to synthesize plausible surrogate representations of absent modalities for training-time augmentation, generative approaches may provide additional support for learning under incomplete multimodal availability, although their benefit depends strongly on how synthetic data are incorporated.

Through these three perspectives, our study aims to advance methodological rigor in the use of ADNI data and to provide new insights into the multimodal characterization of neurodegeneration in aging populations. Beyond their technical relevance, these contributions are intended to foster more equitable, interpretable, and clinically applicable machine learning models for geriatrics.

This work addresses challenges in neuroimaging by proposing a unified framework for studying multimodal classification under incomplete modality availability. It combines LoRA-adapted foundation encoders, specifically DinoV3 [9] and RadioDino [10], with ControlNet-based modality translation and structured subgroup-level evaluation.

The manuscript is organized as follows. We begin with a review of the state of the art related to the investigated topics in Section 2. We then outline the datasets, materials, and methodologies employed in this study in Section 3, followed by a detailed description of the proposed approach in Section 4.

Next, the experimental evaluation is presented in Section 5, which is further divided into the preliminary results in Section 5.2 and the subsequent analyses leading up to the clinical integration discussion in Section 5.7.

A fairness assessment is provided in Section 6, after which we discuss the limitations of the work and potential future directions in Section 7. The manuscript concludes in Section 8, where we summarize the insights gained through our efforts to address the gap in multimodal neurodegenerative disease identification (Fig. 1).

2. Related work

The study of neurodegenerative diseases, particularly AD, has greatly benefited from extensive research efforts integrating neuroimaging, clinical, and demographic data. This section reviews the foundational resources, methodological challenges, and advances in this domain. We begin with an overview of the Alzheimer's Disease Neuroimaging Initiative and its pivotal role in enabling large-scale multimodal analyses and clinical trial biomarker development (Section 2.1). We then discuss fairness issues arising from demographic and acquisition biases in neuroimaging applications, emphasizing their impact on the generalizability of predictive models in Section 2.2. Next, Section 2.3 provides a survey of unimodal neuroimaging approaches, highlighting their contributions and limitations when considered in isolation. Section 2.4 covers multimodal neuroimaging frameworks, which strive for more comprehensive disease representations by integrating multiple imaging and clinical data sources. Finally, in Section 2.5 we address the common challenge of missing data in neuroimaging studies and review advanced imputation and generative modeling techniques aimed at mitigating its adverse effects.

2.1. ADNI and neurodegenerative disease analysis

The Alzheimer's Disease Neuroimaging Initiative has established itself as one of the most influential resources for studying neurodegenerative diseases in aging populations. By providing large-scale longitudinal data that include structural MRI, PET, DTI, and extensive clinical and cognitive assessments, ADNI enables the development of models that integrate anatomical, functional, and clinical perspectives of disease progression [11–13]. Its standardized protocols across multiple sites have further contributed to reproducibility and broad adoption in computational neuroimaging research. While protocol harmonization has been a central objective of the ADNI initiative, substantial acquisition

heterogeneity persists across phases. Earlier releases, such as ADNI1, allowed greater flexibility in scanner hardware and field strength, including both 1.5T and 3T T1-weighted MRI acquisitions, whereas later phases, such as ADNI3, restricted new acquisitions to 3T scanners. As a result, the aggregated ADNI dataset exhibits non-negligible variability in scanner characteristics and acquisition protocols. This heterogeneity more closely reflects real-world clinical imaging conditions and underscores the importance of developing models that remain robust across differences in scanner field strength and acquisition settings.

A wide range of studies have relied on ADNI for classification tasks, including distinguishing between healthy controls, mild cognitive impairment, and Alzheimer's disease, as well as for predicting conversion risk and monitoring progression [14,15]. These applications have highlighted the utility of ADNI in benchmarking machine learning pipelines for geriatrics, serving as a reference point for both unimodal and multimodal approaches. The dataset has also fostered comparisons between imaging modalities, allowing researchers to quantify the relative contributions of MRI, PET, DTI, and clinical information.

Despite these advances, several limitations of ADNI remain a recurring theme in the literature. Demographic and site-specific imbalances can introduce biases into predictive models, limiting their generalizability. In addition, the heterogeneous availability of modalities leads to incomplete datasets that complicate downstream analyses. These challenges underline the need for fairness-aware modeling, robust multimodal fusion, and principled strategies for handling missing data, which together form the core focus of the present study [16].

2.2. Fairness in neuroimaging and ADNI classification

Fairness has become an increasingly important consideration in the application of machine learning to medical imaging, as demographic and acquisition-related biases can lead to unequal model performance across patient groups. In the context of neuroimaging, these biases may stem from factors such as age, sex, education level, or site-specific imaging protocols, all of which can systematically influence predictions and limit the clinical reliability of computational tools [17,18].

Within ADNI, fairness concerns are particularly relevant due to the dataset's heterogeneous composition. Studies have shown that imbalances in sample sizes across demographic subgroups and imaging sites can result in models that are highly accurate on the dominant cohorts but less reliable for underrepresented populations. Moreover, the prevalence of scanner-specific artifacts and acquisition differences raises concerns about domain shifts that may confound disease-related signals with technical variability [19].

Addressing these issues has motivated the development of strategies such as balanced sampling, domain adaptation, bias-regularized training objectives, and subgroup performance evaluation. While such approaches have improved robustness, fairness considerations remain underexplored in ADNI-based classification tasks. In this work, we prioritize fairness as a central element of our analysis, systematically examining subgroup performance and integrating mitigation strategies to ensure that predictive models for neurodegeneration are both accurate and equitable across geriatric populations [20,21].

2.3. Unimodal neuroimaging analysis

Unimodal neuroimaging analysis has been the foundation of many computational approaches to neurodegenerative disease, with studies often focusing on a single modality to simplify modeling and interpretation. Structural MRI has been the most widely used, providing detailed anatomical information on brain atrophy patterns associated with Alzheimer's disease and related conditions. Similarly, PET has enabled the quantification of amyloid and tau deposition [22,23], while DTI has been employed to capture microstructural alterations in white matter pathways.

These unimodal approaches have yielded important insights into the mechanisms and progression of neurodegeneration. MRI-based classifiers, for instance, have consistently demonstrated strong performance in distinguishing between healthy controls, mild cognitive impairment, and Alzheimer's disease [24,25]. PET imaging has been particularly valuable in identifying early pathological changes [26], while DTI has shed light on alterations in connectivity that complement structural and functional findings [27,28]. Each modality thus offers a unique window into the disease process, providing clinically relevant information on its own.

Nevertheless, unimodal analyses face significant limitations. Single modalities often capture only one aspect of disease biology, leaving important complementary information unexploited. Furthermore, unimodal models may be more sensitive to confounding factors such as scanner variability or preprocessing pipelines, reducing generalizability across cohorts. These challenges highlight the importance of moving beyond unimodal strategies to more integrative approaches, while still recognizing the foundational role that unimodal neuroimaging analysis has played in shaping our understanding of neurodegenerative disease.

2.4. Multimodal neuroimaging analysis

Building on unimodal foundations, multimodal neuroimaging analysis has emerged as a powerful strategy to capture complementary aspects of neurodegeneration. By integrating structural MRI, PET, DTI, and clinical information, multimodal approaches aim to provide a more comprehensive representation of disease processes. Early studies combined imaging and non-imaging variables through simple concatenation, while more recent work has explored advanced fusion techniques, including multi-branch neural networks, attention-based mechanisms, and joint embedding frameworks [29–31].

Evidence from these studies suggests that multimodal models often outperform unimodal counterparts in diagnostic and prognostic tasks. For example, combining MRI with PET has been shown to improve the detection of early pathological changes [32]. Clinical and demographic variables further enhance predictive performance by anchoring imaging findings to patient-level context. Together, these results demonstrate the synergistic potential of multimodal learning in geriatrics [33].

Despite these advantages, multimodal neuroimaging remains methodologically challenging. Data incompleteness is common, as not all patients undergo every imaging modality, leading to reduced sample sizes when strict inclusion criteria are applied. Fusion strategies can also suffer from overfitting or modality imbalance, in which one source dominates the predictions at the expense of others. Addressing these limitations requires robust techniques for handling missing data and more principled integration strategies. In this study, we systematically evaluate multimodal and unimodal pipelines, highlighting both their strengths and weaknesses in the context of ADNI classification [34,35].

2.5. Missing data in neuroimaging

Missing data is a pervasive challenge in large-scale neuroimaging studies, including ADNI, where not all participants undergo every imaging modality or complete all clinical assessments. Such gaps can reduce statistical power, introduce bias, and complicate downstream analyses, particularly in multimodal pipelines that rely on complete information from multiple sources.

Traditional approaches to handling missing data have included simple imputation methods such as mean substitution or regression-based prediction, as well as more sophisticated matrix completion techniques. While these strategies can partially mitigate the problem, they often fail to capture the complex relationships between modalities and may introduce additional bias if the missingness is non-random. While several works aim to maximize mutual information [36–38] more recently, deep generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion-based models, have been applied to learn plausible representations of missing modalities by

leveraging correlations across available data [34,39,40]. In the broader medical imaging domain, approaches like HeMIS [41] have also been proposed, which create modality-specific embeddings that are combined via summary statistics, providing robustness to missing modalities without requiring explicit imputation. For health records, which are a core data for diagnostics methods like M3Care [42] has been proposed to overcome generative approaches related problems. Despite their promise, generative approaches also present challenges. Model training can be computationally intensive, and ensuring that synthetic or imputed data accurately reflect true biological signals remains non-trivial. Nevertheless, these methods offer a principled framework for addressing missing data, enabling more complete and robust analyses in neurodegenerative disease research. In this work, we investigate generative models as a training-time augmentation strategy to support multimodal learning under missing-modality conditions, rather than as direct replacements for unavailable clinical acquisitions.

3. Materials and methods

This section outlines the materials and methods employed in the proposed research and is organized as follows: foundation models used for multimodal neuroimaging analysis are introduced in Section 3.1, the adaptation strategy based on Low-Rank Adaptation is detailed in Section 3.2, the generative modality translation pipeline using ControlNet is presented in Section 3.3, and the characteristics of the ADNI dataset are described in Section 3.4.

3.1. Foundation models

In this work, foundation models were leveraged as the backbone for the analysis of neuroimaging data.

Owing to their ability to learn generalizable representations from large-scale datasets, these models provide a robust starting point for downstream tasks in neurodegenerative research.

Pre-trained vision transformers were employed to extract imaging biomarkers from DTI, MRI, and PET data. The models were subsequently fine-tuned for classification tasks to distinguish between Alzheimer's disease patients, individuals with mild cognitive impairment, and cognitively normal controls in a multiclass (MC) setting, or between Alzheimer's disease patients and controls in a binary (B) setting.

This approach reduces the reliance on handcrafted features and enables a more comprehensive representation of disease-related patterns. In particular, the use of foundation models supports better generalization across cohorts and mitigates the variability introduced by acquisition protocols, which is critical for reproducibility in Alzheimer's disease detection studies.

3.2. Low-rank adaptation

To efficiently fine-tune large foundation models for Alzheimer's disease detection, Low-Rank Adaptation (LoRA) [43] was employed. LoRA introduces trainable low-rank decomposition matrices into the attention layers of pre-trained transformers, substantially reducing the number of parameters that must be updated during training. Rather than fine-tuning the full weight matrices of the self-attention mechanism, LoRA factorizes these matrices into lower-dimensional components that are learned during adaptation, while the original weights remain frozen. This design allows the model to capture task-specific information with minimal additional parameters, resulting in faster convergence, lower memory requirements, and improved scalability when adapting large foundation models to specialized domains such as neuroimaging.

This approach enables the adaptation of foundation models to neuroimaging data without the need for full retraining, thereby reducing computational costs and mitigating overfitting when working with limited labeled datasets.

In this study, LoRA was applied to the vision transformer backbone, enabling the extraction of Alzheimer's-related imaging biomarkers while preserving the general representations learned during large-scale pre-training. This strategy allowed the models to specialize in downstream classification tasks, distinguishing Alzheimer's disease patients from cognitively normal controls, as well as handling MC settings that included mild cognitive impairment.

3.3. ControlNet for modality translation

To address the heterogeneity of neuroimaging data and reduce the impact of missing modalities, ControlNet [44] was incorporated into the pipeline. ControlNet extends pre-trained diffusion models by conditioning the generative process on structural priors, such as anatomical or modality-specific features, which enable controlled image-to-image translation between neuroimaging modalities. In this setting, the model receives guidance from structural information extracted from the source modality while learning to generate a realistic representation in the target modality. This mechanism ensures that anatomical consistency is preserved across translations, while modality-specific contrasts are adapted to reflect the imaging characteristics of the target domain. By explicitly disentangling structural and modality information, ControlNet provides a robust framework for synthesizing missing modalities, harmonizing heterogeneous datasets, and facilitating downstream analyses in multimodal neuroimaging studies.

In this study, ControlNet was employed to translate between DTI, MRI, and PET. This translation allows the synthesis of complementary imaging data when a modality is unavailable, thereby mitigating data sparsity and ensuring more consistent multimodal representations across patients.

By generating modality-consistent surrogate neuroimaging data, ControlNet provides a mechanism for studying training-time augmentation under missing-modality conditions. In the present work, its downstream effect is configuration-dependent: in some settings it supports multimodal learning, whereas in others it introduces harmful variability and degrades performance.

3.4. ADNI dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD.

The original goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and providing comprehensive data concerning the diagnosis and progression of Alzheimer's disease to the scientific community.

ADNI has evolved through multiple phases (ADNI1, ADNI-GO, ADNI2, ADNI3, and the current ADNI4), continuously expanding its cohort and incorporating state-of-the-art imaging techniques, biomarker assays, and cognitive assessments. The current ADNI4 phase focuses on recruiting a more diverse participant population, increasing the use of advanced imaging modalities, and collecting longitudinal data to improve understanding of Alzheimer's disease progression across different demographic and clinical subgroups.

The comprehensive and publicly available nature of ADNI data has made it a cornerstone resource for developing and validating neuroimaging biomarkers, machine learning models, and other computational approaches for Alzheimer's disease research.

All experiments were conducted using ADNI data downloaded from the LONI Image Data Archive, including ADNI1, ADNI-GO, ADNI2, ADNI3, and ADNI4. Data were accessed in August 2025. Patient filtering (first visit only, consistent diagnosis across visits) and all preprocessing

steps were implemented using publicly available scripts provided in our code repository, enabling exact replication of cohort construction.

4. Proposed approach

This section presents the proposed approach for multimodal neurodegenerative disease analysis, outlining a systematic workflow designed to maximize reproducibility and diagnostic accuracy. Data preprocessing steps, including cohort filtering and slice generation, are detailed in Section 4.1. The establishment of baseline models to evaluate single-slice and single-modality performance is described in Section 4.2. Subsequent multimodal integration experiments and fusion strategies are presented in Section 4.3, followed by the generative augmentation method used to reconstruct missing modalities in Section 4.4. The incorporation of clinical and demographic features is discussed in Section 4.5, while the final multimodal fusion heads are outlined in Section 4.6. Collectively, these components form a unified framework for studying multimodal Alzheimer's disease classification under incomplete modality availability.

4.1. Data processing

Our proposed approach systematically addresses multiple concerns that can compromise experimental validity and lead to artificially inflated performance due to the so-called data leakage phenomenon [5].

Specifically, we aim to achieve two key objectives: (i) provide a blueprint for reproducible results, and (ii) deliver a comprehensive and efficient multimodal, foundation model-based solution for neurodegenerative disease analysis.

For the first objective, we collect all patients from the ADNI database up to ADNI4, the most recent release. We then select only the most common preprocessing techniques applied to the corresponding patient data.

No additional preprocessing was applied to MRI or DTI data beyond what was provided by ADNI. PET slice-level representations were already preprocessed by ADNI; for our experiments, we combined these slices into full 3D volumes and subsequently resliced them into the three orthogonal planes. All slices are obtained by using the middle slice of the selected axis. This middle-slice formulation was adopted primarily to preserve compatibility with the pretrained 2D foundation encoders used in this study and to maintain a unified experimental protocol across modalities, fusion strategies, and missing-modality settings. At the same time, it discards through-plane spatial correlations and may miss volumetric patterns relevant to Alzheimer's disease.

Next, we filter the cohort to include only first-visit scans, in order to design a screening-oriented solution rather than focusing on late-stage diagnosis, which may be less informative due to already extensive brain tissue damage. A further crucial filtering step involves excluding patients whose diagnosis changes across visits, as well as cases labeled as uncertain, to ensure the reliability of diagnostic information. This filtering step removed 423 patients, corresponding to approximately 18.7% of the total cohort. We acknowledge that this criterion excludes clinically relevant cases such as MCI-to-AD converters and may slightly alter the underlying demographic and clinical distributions. After applying these filters, the final merged dataset comprises 1841 unique subjects across modalities.

The final step for imaging data differs across modalities. For MRI and DTI data, we generate axial, sagittal, and coronal slices. For PET

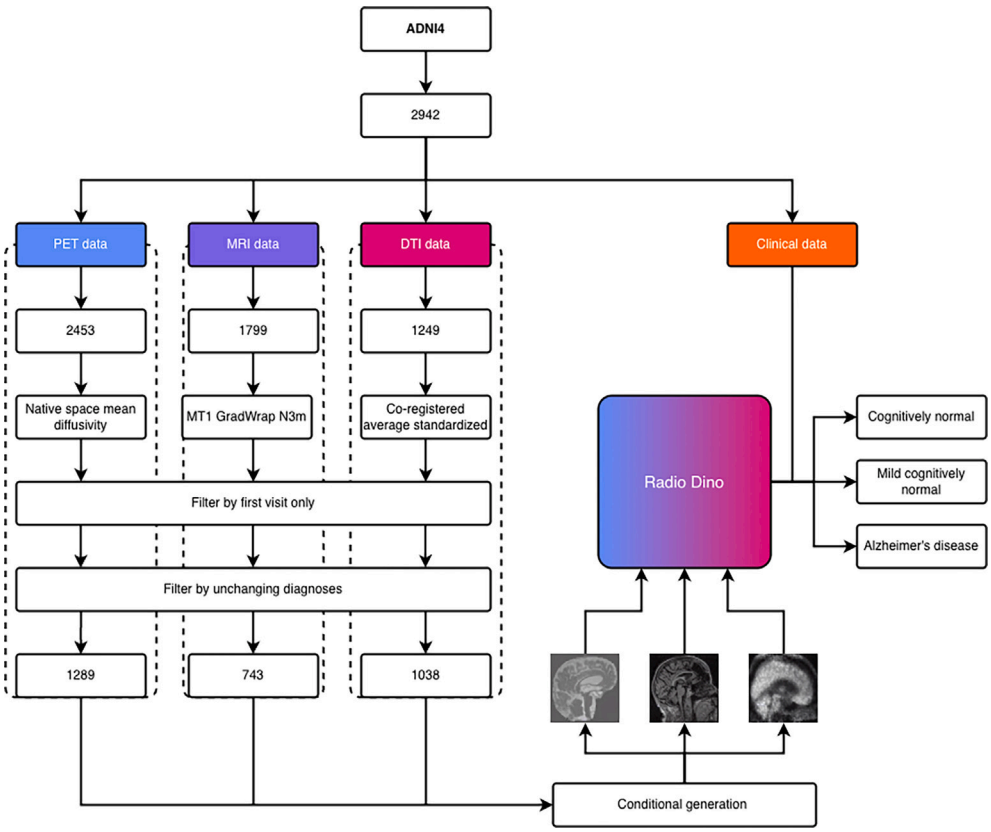


Fig. 2. An overview of the patient selection and data preparation workflow. This diagram outlines the filtering of the ADNI cohort to first-visit scans with consistent diagnoses. It then details the modality-specific image processing, generating axial, sagittal, and coronal slices for MRI and DTI, alongside reslicing for PET scans, which creates a standardized dataset for our multimodal analysis framework.

Table 1
Overview of multimodal experiments. A checkmark (✓) indicates the inclusion of the corresponding modality–slice configuration.

Experiment	Axial MRI	Sagittal MRI	Coronal MRI	Axial DTI	Sagittal DTI	Coronal DTI	Axial PET	Sagittal PET	Coronal PET
MRI-Full	✓	✓	✓						
DTI-Full				✓	✓	✓			
PET-Full							✓	✓	✓
Axial-Cross	✓			✓			✓		
Sagittal-Cross		✓			✓			✓	
Coronal-Cross			✓			✓			✓
MRI-DTI	✓	✓	✓	✓	✓	✓			
MRI-PET	✓	✓	✓				✓	✓	✓
DTI-PET				✓	✓	✓	✓	✓	✓
Full	✓	✓	✓	✓	✓	✓	✓	✓	✓

data, since slice-level representations are already available, we first reconstruct a full 3D volume for each patient and subsequently reslice it into the three orthogonal planes. We report an overview schema of our approach in Fig. 2.

4.2. Baseline establishment

In order to investigate whether neurodegenerative analysis benefits from multimodal approaches, we first conduct a set of baseline experiments to evaluate performance across different levels of data granularity. Specifically, we train models on imaging data while also separating samples by slice type. This allows us to assess whether a single slice is sufficient for analysis, or whether incorporating multiple views improves classification performance.

In our approach, we leverage two foundation models: DinoV3 [9], a novel generalist model pretrained on a broad dataset, and RadioDINO [10], a DINO-inspired foundation model pretrained on a large-scale radiological dataset [45]. Convolutional-based foundation models are not included, in order to apply the same Low-Rank Adaptation parameters across all models, thereby enhancing the fairness and reproducibility of our experiments.

Furthermore, we perform single-slice, single-modality classification experiments in two distinct scenarios: (i) CN vs. MCI vs. AD, namely MC, to enable fine-grained analysis, and (ii) CN vs. AD, namely B, to focus on a more specific diagnostic setting.

4.3. Multimodal analysis

To comprehensively evaluate the impact of multimodal integration, we design experiments that combine information at different levels of granularity. In particular, we consider configurations based on individual slice orientations for each modality (e.g., axial PET, sagittal MRI), as well as combinations of modalities with all slices. This allows us to disentangle the contribution of single-slice, single-modality inputs from multimodal and multislice fusion. The complete set of configurations is summarized in Table 1. From this point onward, we will refer to each configuration using the experiment names provided in Table 1.

4.4. Generative augmented classification

To address the problem of missing or incomplete multimodal data, we introduce a generative augmentation strategy aimed at reconstructing absent imaging modalities from the available ones. This approach enables a consistent multimodal representation across patients and prevents the exclusion of cases lacking complete imaging information.

We follow the assumption that each patient has at least one available imaging modality. Based on this premise, we employ a modality translation framework capable of generating synthetic counterparts of the missing modalities through conditional diffusion models. Specifically, we adopt the ControlNet architecture [44], which extends diffusion-based generative models by conditioning the denoising process on structural information extracted from a source modality. This conditioning mechanism enforces anatomical consistency while adapting image contrast and appearance to match the target modality. In our implementation, we use the standard HuggingFace diffusers ControlNet configuration, where conditioning is injected through the default ControlNet residual branches of the U-Net downsampling and mid blocks, without custom layer-wise restriction. The conditioning scale was fixed to 1.0 throughout the modality-translation experiments. Because this pipeline requires a text input, we supplied the same fixed prompt, namely "-", for every sample during both training and inference. No diagnosis, demographic, or patient-specific information was ever used as text conditioning. As a result, no label information could leak through the text branch, and all effective conditioning originated exclusively from the source-modality image.

During training, each modality pair is modeled independently, with the source modality serving as the input and the target modality as the supervision signal. The model learns a mapping between corresponding anatomical regions while preserving spatial alignment.

The full objective function for ControlNet training is given by:

$$\mathcal{L} = \mathbb{E}_{x,t,\epsilon}\Big[\underbrace{\|\epsilon_\theta(z_t, t, y, y_{\text{cond}}) - \epsilon\|_2^2}_{\text{reconstruction / denoising loss}}\Big], \tag{1}$$

where x denotes the input image from the target modality, y is the corresponding ground-truth target image, y_{cond} represents the source modality used for conditioning, t is a diffusion timestep randomly sampled from $[0, T]$, z_t is the noisy latent representation of y at timestep t obtained via the forward diffusion process, ϵ is the Gaussian noise added to the latent to produce z_t, and $\epsilon_\theta(z_t, t, y, y_{\text{cond}})$ is the network's predicted noise. The loss $\mathcal{L}$ corresponds to the mean squared error between the predicted and actual noise, while the conditioning information y_{cond} is enforced directly through the ControlNet architecture rather than as an explicit additional loss term.

The ControlNet model was trained for 20,000 iterations using a batch size of 32, resolution of 224×224, AdamW optimizer, and a weight decay of 0.01. A constant learning rate scheduler with a linear warmup of 500 steps was employed, and the initial learning rate was set to 5×10^{-6}. Parameters were selected after a preliminary qualitative analysis. Images were normalized using the standard ImageNet mean and standard deviation to maintain compatibility with the pretrained ControlNet weights, as is common in transfer learning with pretrained diffusion models. No additional validation of these statistics on medical images was performed. All images were additionally partially aligned such that the brain occupies a consistent field of view across subjects.

Once trained, the diffusion-based translation models generate synthetic modalities for patients missing one or more imaging types.

These generated representations are then integrated into the downstream classification pipeline as additional inputs, effectively expanding the dataset and enabling multimodal fusion even for incomplete patient profiles. To further investigate the degradation observed under synthetic

augmentation, we introduced two modifications to the training procedure in an additional ablation. First, we replaced the previous uniform classification objective with a weighted synthetic-loss formulation,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{real}} + \lambda_{\text{syn}}\mathcal{L}_{\text{synth}},$$

where $\mathcal{L}_{\text{real}}$ denotes the loss computed on real inputs, $\mathcal{L}_{\text{synth}}$ denotes the loss computed on synthetic inputs, and $\lambda_{\text{syn}} < 1$ reduces the contribution of synthetic samples during optimization. Second, we made the synthetic-data usage policy explicit. Under a *non-selective* policy, synthetic modalities are introduced regardless of whether the corresponding real modality is already available. Under a *missing-only* policy, synthetic modalities are used only when the corresponding real modality is unavailable. This allowed us to distinguish the effect of reducing the optimization weight of synthetic samples from the effect of restricting their use to genuinely missing modalities.

Within the proposed framework, generative augmentation is used as a training-time strategy to expand multimodal supervision under incomplete-modality conditions. Its intended role is not to replace real acquisitions, but to provide auxiliary synthetic inputs that may support learning when real modalities are unavailable.

All data used in this study were obtained under the ADNI Data Use Agreement and consist exclusively of de-identified, publicly available research data. The generation of synthetic images is employed solely as an internal training augmentation strategy to address missing imaging modalities and does not involve redistribution of original ADNI data or derivative datasets. Synthetic images are not released or shared; in addition, pretrained generative models capable of producing ADNI-derived images are not distributed.

4.5. Clinical and demographic integration

Beyond imaging information, we incorporate non-imaging clinical data to enhance the model's ability to capture individual patient variability and disease-related factors that may not be directly observable in imaging modalities. Specifically, we leverage two key sources of tabular information from the ADNI database: the PTDEMOG and AMAS questionnaires.

The PTDEMOG form includes demographic and baseline information such as age, gender, handedness, and education level. The AMAS questionnaire provides multiple cognitive subscales assessing general knowledge, memory, attention, language, and executive function. In this study, all available subscales and demographic variables were included to maximize the information captured, except for the PTADBEG, PTCOGBEG, and PTADDX variables, which were excluded to prevent data leakage, as their values are directly related to the year of AD or MCI diagnosis. Data were preprocessed to retain only the first visit per patient, with missing values imputed as -4, similar to missing values in other related ADNI datasheets, and categorical entries split at the '|' character to ensure a single value per feature. No samples were excluded due to missing clinical data. This preprocessing ensures that all demographic and cognitive subscales can be incorporated into the projection head, maximizing the information provided to the model. In this study, clinical and demographic features from PTDEMOG and AMAS were used in their raw form without explicit scaling or standardization. We note that caution is needed when generalizing to external cohorts with different feature distributions.

To process these structured features, we design a lightweight projection head that maps tabular data into the same latent space as the imaging representations. Let $\mathbf{x} \in \mathbb{R}^F$ denote the vector of clinical features for a given subject, where $F = \texttt{num\_features}$. Clinical information is incorporated via a projection head that maps $\mathbf{x}$ into the same embedding space as the imaging representations, with dimensionality d_{model}. This mapping is implemented as a sequence of linear transformations with normalization and non-linear activations:

$$\begin{aligned}
\mathbf{h}_1 &= \phi\big(\mathbf{W}_1\,\mathrm{LN}(\mathbf{x}) + \mathbf{b}_1\big), \quad & \mathbf{W}_1 &\in \mathbb{R}^{d_{\text{model}}\times F},\\
\mathbf{h}_2 &= \phi\big(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2\big), & \mathbf{W}_2 &\in \mathbb{R}^{4d_{\text{model}}\times d_{\text{model}}},\\
\mathbf{z}_{\text{clin}} &= \mathbf{W}_3\,\mathrm{LN}(\mathbf{h}_2) + \mathbf{b}_3, & \mathbf{W}_3 &\in \mathbb{R}^{d_{\text{model}}\times 4d_{\text{model}}},
\end{aligned} \tag{2}$$

where $\phi(\cdot)$ denotes the GELU activation and LN indicates layer normalization. Dropout with probability 0.1 is applied to the input features prior to the first linear transformation. The intermediate expansion to $4d_{\text{model}}$ follows a standard dimensional scaling commonly used in transformer feed-forward blocks and is adopted here as a practical architectural choice to increase intermediate capacity before projection back to d_{model}, without implying any explicit feature disentanglement property. The resulting embedding $z_{\text{clin}} \in \mathbb{R}^{d_{\text{model}}}$ is aligned with the dimensionality of imaging-derived embeddings and can be incorporated as an additional token in the multimodal fusion stage. This design enables the joint integration of imaging and clinical information within a unified embedding space and provides a modular interface for incorporating additional structured variables when available.

This integration strategy allows the network to jointly reason over structural, metabolic, and demographic cues, capturing a more comprehensive picture of neurodegenerative progression. It also serves as a flexible interface for incorporating additional questionnaire-based features in future extensions of the framework.

4.6. Fusion heads

The final multimodal representation integrates embeddings from all enabled imaging encoders together with the clinical projection obtained from tabular data. To derive the diagnostic output, we explore two distinct fusion heads that differ in how they perform cross-modal fusion and information aggregation.

- **Mean Head.** In this configuration, the model computes a simple average of the `CLS` tokens extracted from each imaging encoder together with the embedding corresponding to the tabular clinical data. Formally, for embeddings $X_i \in \mathbb{R}^{B\times d_{\text{model}}}$ from each modality i, the fused embedding is computed as

  $$X_{\text{fused}} = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

  where N is the number of enabled modalities plus clinical embeddings. This approach assumes that relevant diagnostic information is distributed across modalities, promoting robustness and interpretability through global token averaging. It introduces no additional learnable parameters.
- **Transformer Head with CLS Token.** In this configuration, the set of imaging and clinical embeddings is processed by a dedicated TransformerEncoder composed of six encoder layers. Each layer consists of a multi-head self-attention mechanism with 8 attention heads, a hidden dimension equal to d_{model}, and dropout applied at 0.1. A learnable CLS token is prepended to the input sequence. Learnable positional embeddings are added to all tokens, and attention masking is applied such that the CLS token can attend to all modality tokens, enabling the model to learn a global, attention-based representation for classification. The transformer output corresponding to the CLS token is passed to a final linear layer for prediction. Formally, given an input sequence of embeddings $X \in \mathbb{R}^{B\times N\times d_{\text{model}}}$, where B is the batch size and N the number of modalities plus the CLS token, each Transformer layer computes:

  $$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V$$

  $$X' = \mathrm{FFN}(\mathrm{LayerNorm}(X + \mathrm{Attention}(X,X,X)))$$

  where $d_k = d_{\text{model}}/\text{num\_heads}$ and FFN denotes a two-layer feed-forward network with GELU activations. The six-layer transformer

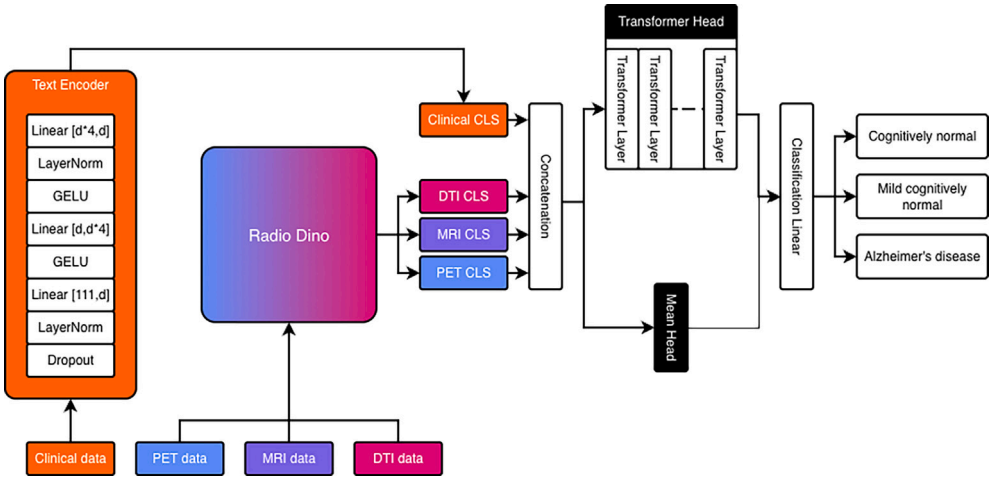


Fig. 3. Proposed multimodal architecture combining clinical data and neuroimaging modalities, PET, MRI, and DTI through foundation model encoding. Clinical features are processed via a separate encoder, while imaging modalities generate specialized CLS tokens. The architecture supports two fusion strategies: a mean head for simple aggregation and a transformer-based fusion, depicted in black boxes with white text, culminating in MC or B classification.

fusion head contains approximately 33.1M parameters, which is larger than smaller vision transformer modules such as 22M in RadioDino small, but was chosen to balance the combined multimodal input capacity with classification performance.

Both heads share a common final linear classification layer projecting the resulting embedding into the disease label space. This dual-head design enables a fair comparison between simple aggregation mechanisms and attention-based reasoning, offering insight into how different fusion strategies influence multimodal interpretability and generalization.

This allows the framework to jointly reason over structural, metabolic, and demographic cues, capturing a holistic view of neurodegenerative disease progression. We report a full depiction of the whole architecture in Fig. 3

Relation to established multimodal fusion strategies. The fusion strategies evaluated in this work were deliberately selected to represent two complementary and widely adopted paradigms in multimodal learning: parameter-free aggregation and learnable attention-based integration. Simple linear aggregation of modality-specific representations has long been used as a baseline in multimodal settings and is closely related to classical multiview learning approaches, including canonical correlation analysis, which aim to combine information from multiple modalities through shared or correlated latent representations. By averaging modality-specific embeddings extracted by pretrained encoders in a shared feature space, the CLS mean head provides a correlation-agnostic yet computationally efficient fusion mechanism that has been widely adopted as a strong baseline in prior multimodal studies. Conversely, the Transformer-based fusion head explicitly models cross-modal interactions through self-attention, aligning with more recent attention-driven multimodal fusion architectures that learn modality interactions in a data-driven manner. Rather than reimplementing task-specific multimodal fusion networks, which often rely on handcrafted features or modality-specific optimization schemes, our framework evaluates these two representative fusion extremes within a unified foundation model setting. This design enables a controlled comparison that isolates the impact of fusion complexity while maintaining consistent encoders, training procedures, and evaluation protocols across all experiments. In the present framework, multimodal fusion is performed over a unified token sequence rather than through modality-specific cross-attention branches.

5. Experimental evaluation

In this section, we present a comprehensive experimental evaluation of our proposed multimodal framework for Alzheimer's disease characterization. We begin by describing our training protocol and hyperparameters, followed by baseline results using single-modality approaches across different neuroimaging modalities (MRI, DTI, and PET). We then evaluate our multimodal fusion strategies, comparing mean-based aggregation with transformer-based heads across various combinations of modalities and slice orientations. Subsequently, we assess the effectiveness of our generative augmentation approach for handling missing data, demonstrating how synthetic data generation can enhance classification performance. Finally, we integrate clinical data into our framework and analyze the contribution of each data integration to the overall diagnostic accuracy. All experiments are conducted using rigorous cross-validation protocols to ensure robust and reproducible results. This section begins by detailing the classification training protocol and hyperparameters in Section 5.1, followed by the training strategy selection and preliminary ablation study in Section 5.2. Baseline results using single-modality approaches are reported in Section 5.3, and multimodal fusion performance is evaluated in Section 5.5. The impact of generative augmentation for handling missing modalities is analyzed in Section 5.6, and finally, the integration of clinical and demographic data within the generative framework is assessed in Section 5.7.

5.1. Classification training protocol and hyperparameters

All experiments were implemented in PyTorch and executed on GPU, specifically an on-demand A100 80GB for the generation experiments for approximately 70 GPU-hours. For all other experiments we leveraged an in-house workstation with a consumer-grade 4060ti 16GB GPU, for approximately 40 GPU-days. We employed a repeated stratified group k-fold cross-validation scheme with $k = 5$ folds and $n = 5$ repetitions, resulting in 25 model fits per configuration. To ensure robustness and reproducibility, seeds were initialized at 42 and incremented by 1 for each repetition. For model evaluation, subjects were first split into training and test sets using a grouped splitting strategy based on subject identifiers. Within the training set, repeated stratified group k-fold cross-validation was applied, using the subject identifier as the grouping key and the subject-level diagnosis for stratification. All slices and visits belonging to the same subject were assigned to the same fold, preventing subject-level information leakage while maintaining class balance across folds. While deterministic, this choice ensures controlled variability across repetitions and allows exact replication of the

experimental protocol. The backbone models were Vision Transformers from the HuggingFace Hub, with input images resized to 224×224 pixels and normalized using standard ImageNet statistics. Optimization was performed using AdamW with a learning rate of 3×10^{-4} and a cosine annealing schedule. The batch size was set to 32, and each fold was trained for 30 epochs using cross-entropy loss. These hyperparameters were selected based on a combination of prior literature on similar multimodal neuroimaging tasks [46] and preliminary experimentation that indicated stable convergence and satisfactory performance. For the LoRA [43] parameters, we used the original PEFT implementation with a dropout of 0.1, rank of 8, and an alpha of 8, while we selected "all-linear" as the target layer selection method. Gradient clipping with an ℓ_2 norm of 3.0 was applied, and mixed-precision training was optionally enabled to accelerate training while reducing memory usage. For statistical testing, we report p_a values computed using all 25 runs, across all folds and repetitions, as the sample set, whereas p_i values are reported using only $N = 5$, corresponding to the number of repetitions, while ignoring the individual folds. This distinction clarifies whether the full set of model fits or only the independent repetitions are considered for significance estimation. To ensure a fairer comparison with volumetric transformer baselines, both 3D transformer models were retrained using the same transformer-oriented optimization setup: AdamW optimizer, initial learning rate 5×10^{-5}, weight decay 1×10^{-4}, linear warmup followed by cosine decay with warmup over the first 10% of training, minimum learning rate 1×10^{-6}, batch size 2, gradient accumulation 8, 100 training epochs, and gradient clipping with maximum norm 1.0. In the revised volumetric comparison, all 3D backbones were implemented using standard MONAI architectures. ResNet18 and ResNet50 were initialized from MedicalNet pretraining [47], while the Swin baseline was initialized from pretrained Swin UNETR encoder weights available within the MONAI ecosystem [48]. As these Swin weights correspond to the encoder only, the downstream classification layers were learned during fine-tuning. For DenseNet121 and the plain ViT baseline, we did not identify official MONAI-distributed pretrained volumetric classification weights matching our setting to the best of our knowledge, and these models were therefore trained using the standard MONAI initialization. Classification-related hyperparameters are reported in this subsection, while generation-specific training settings are provided separately in Section 5.6.

Table 2
Ablation study on the impact of adaptation strategy for DTI axial slice classification, comparing LoRA, full fine-tuning, and head-only fine-tuning across DinoV3 and RadioDino models. LoRA consistently yields the highest and most stable F1 scores in both MC and B scenarios, motivating its use as the default adaptation strategy in all subsequent multimodal and multislice experiments.

Slice	Model / Strategy	F1 (MC) (%)	F1 (B) (%)
Modality: DTI			
Axial	DinoV3 base (LoRA)	**39.08 ± 4.69**	**60.76 ± 5.68**
	DinoV3 base (full finetune)	34.21 ± 5.12	51.94 ± 6.15
	DinoV3 base (head finetune)	34.87 ± 5.05	52.63 ± 6.01
	DinoV3 small (LoRA)	**39.76 ± 6.79**	**61.53 ± 6.24**
	DinoV3 small (full finetune)	34.58 ± 7.21	52.10 ± 6.83
	DinoV3 small (head finetune)	35.12 ± 7.14	52.82 ± 6.76
Modality: DTI			
Axial	RadioDino base (LoRA)	**37.62 ± 5.42**	**58.56 ± 7.08**
	RadioDino base (full finetune)	34.12 ± 5.91	52.03 ± 7.52
	RadioDino base (head finetune)	34.76 ± 5.86	52.75 ± 7.43
	RadioDino small (LoRA)	**36.53 ± 5.18**	**59.44 ± 8.25**
	RadioDino small (full finetune)	34.01 ± 5.63	52.20 ± 8.79
	RadioDino small (head finetune)	34.65 ± 5.57	52.92 ± 8.68

5.2. Training strategy selection and preliminary ablation study

Our first step was to systematically determine the most robust and generalizable training strategy through an ablation study, using DTI axial as our reference setting since it was the first and simplest single-slice experiment conducted. Specifically, we compared LoRA adaptation, full finetuning, and head-only finetuning schemes across both DinoV3 and RadioDino model variants. As summarized in Table 2, LoRA adaptation yielded the most consistent advantage. For instance, on the DTI axial dataset, the LoRA-adapted DinoV3 base achieved 39.08 ± 4.69 (MC) and 60.76 ± 5.68 (B) mean F1, outperforming full finetuning (34.21 ± 5.12 / 51.94±6.15) and head-only finetuning (34.87±5.05 / 52.63±6.01). Similar patterns emerged for the RadioDino base (LoRA: 37.62 ± 5.42 / 58.56 ± 7.08), confirming that LoRA-based adaptation not only stabilizes training but also boosts performance relative to traditional finetuning strategies. Based on these results, all subsequent multimodal experiments uniformly adopted LoRA as the adaptation protocol.

5.3. Baseline results

Single-modality approaches offer a direct and interpretable baseline for evaluating classification pipelines, facilitated by extensive benchmarking from established datasets. However, relying solely on one modality inevitably constrains the model's ability to capture complementary or synergistic patterns found across imaging types. We benchmarked a range of widely used convolutional architectures across DTI, MRI, and PET, and across axial, coronal, and sagittal planes. Overall,

Table 3
Detailed classification results for single-modality, single-slice experiments across DTI, MRI, and PET. Results (mean ± standard deviation) are reported for MC and B training. Best-performing models per slice and metric are highlighted in bold.

Modality	Slice	Model	F1 (MC) (%)	F1 (B) (%)
DTI	Axial	ConvNeXt small	26.19 ± 6.75	50.31 ± 11.32
		DenseNet-121	39.05 ± 5.67	**63.92 ± 4.66**
		EfficientNet-B0	**40.62 ± 5.23**	62.23 ± 6.39
		ResNet50	33.36 ± 4.91	55.52 ± 6.45
	Coronal	ConvNeXt small	23.55 ± 1.32	46.69 ± 7.94
		DenseNet-121	39.02 ± 5.67	63.27 ± 8.18
		EfficientNet-B0	**39.94 ± 4.75**	**64.67 ± 6.65**
		ResNet50	38.65 ± 5.20	63.67 ± 8.77
	Sagittal	ConvNeXt small	23.65 ± 2.00	44.10 ± 0.78
		DenseNet-121	30.16 ± 4.05	47.46 ± 3.93
		EfficientNet-B0	**32.60 ± 3.79**	**47.82 ± 4.73**
		ResNet50	31.18 ± 3.90	44.19 ± 1.25
MRI	Axial	ConvNeXt small	21.88 ± 6.32	54.07 ± 16.68
		DenseNet-121	42.24 ± 7.11	**73.22 ± 5.75**
		EfficientNet-B0	**42.35 ± 5.32**	64.10 ± 5.05
		ResNet50	41.05 ± 8.00	69.28 ± 5.23
	Coronal	ConvNeXt small	20.96 ± 2.83	50.91 ± 16.07
		DenseNet-121	**41.07 ± 5.16**	64.84 ± 6.65
		EfficientNet-B0	37.89 ± 6.89	64.72 ± 8.37
		ResNet50	37.61 ± 8.02	**65.14 ± 8.33**
	Sagittal	ConvNeXt small	20.58 ± 1.73	43.58 ± 12.32
		DenseNet-121	**39.51 ± 4.78**	63.22 ± 6.74
		EfficientNet-B0	39.40 ± 4.64	**63.31 ± 7.47**
		ResNet50	35.06 ± 6.80	62.61 ± 8.60
PET	Axial	ConvNeXt small	21.94 ± 1.03	43.30 ± 16.02
		DenseNet-121	**36.40 ± 8.75**	**66.53 ± 5.30**
		EfficientNet-B0	33.65 ± 6.37	58.42 ± 6.05
		ResNet50	22.68 ± 2.01	47.54 ± 7.98
	Coronal	ConvNeXt small	21.94 ± 1.03	58.90 ± 23.94
		DenseNet-121	**49.63 ± 5.65**	**81.59 ± 3.40**
		EfficientNet-B0	46.32 ± 6.34	76.47 ± 4.84
		ResNet50	46.84 ± 7.37	78.84 ± 2.90
	Sagittal	ConvNeXt small	21.94 ± 1.03	54.23 ± 24.21
		DenseNet-121	**35.65 ± 6.82**	**68.11 ± 11.76**
		EfficientNet-B0	33.12 ± 4.97	59.03 ± 7.76
		ResNet50	28.05 ± 5.35	57.19 ± 8.71

Table 4
Detailed classification results for single-modality, single-slice experiments across DTI, MRI, and PET. Results (mean ± standard deviation) indicate that while performance was generally similar for DTI and MRI modalities across both RadioDino and DinoV3 models, PET data demonstrated substantially higher F1 scores for RadioDino, especially in B classification with absolute improvements exceeding 8-10% in several scenarios.

Modality	Slice	Model	F1 (MC) (%)	F1 (B) (%)
DTI	Axial	DinoV3 base	39.08 ± 4.69	60.76 ± 5.68
		DinoV3 small	**39.76 ± 6.79**	**61.53 ± 6.24**
		RadioDino base	37.62 ± 5.42	58.56 ± 7.08
		RadioDino small	36.53 ± 5.18	59.44 ± 8.25
	Coronal	DinoV3 base	**39.97 ± 5.19**	**67.67 ± 6.34**
		DinoV3 small	39.35 ± 4.94	66.83 ± 7.37
		RadioDino base	38.28 ± 5.23	64.86 ± 6.52
		RadioDino small	35.10 ± 4.67	60.57 ± 7.05
	Sagittal	DinoV3 base	29.53 ± 5.65	**46.78 ± 5.03**
		DinoV3 small	29.90 ± 4.27	45.79 ± 3.33
		RadioDino base	**30.37 ± 4.45**	45.06 ± 2.95
		RadioDino small	26.63 ± 3.32	44.27 ± 0.97
MRI	Axial	DinoV3 base	**45.93 ± 6.60**	**70.92 ± 7.38**
		DinoV3 small	41.94 ± 7.84	69.50 ± 6.86
		RadioDino base	43.04 ± 4.40	70.60 ± 5.29
		RadioDino small	39.32 ± 7.29	70.11 ± 7.35
	Coronal	DinoV3 base	**41.71 ± 6.05**	65.95 ± 6.94
		DinoV3 small	41.26 ± 7.98	**68.12 ± 5.47**
		RadioDino base	40.72 ± 7.38	65.76 ± 6.40
		RadioDino small	39.56 ± 6.02	65.89 ± 5.02
	Sagittal	DinoV3 base	36.91 ± 7.26	59.33 ± 10.29
		DinoV3 small	**40.41 ± 8.15**	59.55 ± 7.30
		RadioDino base	39.52 ± 7.43	60.16 ± 9.44
		RadioDino small	39.74 ± 8.18	**62.21 ± 8.13**
PET	Axial	DinoV3 base	38.31 ± 6.39	69.91 ± 5.51
		DinoV3 small	35.25 ± 7.08	68.37 ± 6.42
		RadioDino base	**42.09 ± 7.39**	78.75 ± 7.33
		RadioDino small	40.78 ± 6.22	**78.90 ± 5.38**
	Coronal	DinoV3 base	**52.01 ± 4.41**	79.21 ± 2.83
		DinoV3 small	48.63 ± 3.65	76.90 ± 4.57
		RadioDino base	48.40 ± 5.16	**81.73 ± 3.71**
		RadioDino small	41.24 ± 4.42	77.76 ± 4.68
	Sagittal	DinoV3 base	45.00 ± 5.85	77.38 ± 4.50
		DinoV3 small	42.18 ± 5.30	76.70 ± 3.33
		RadioDino base	**45.83 ± 4.75**	**78.63 ± 3.71**
		RadioDino small	40.75 ± 6.10	77.82 ± 5.13

performance was strongly dependent on both modality and anatomical orientation, with no single architecture consistently dominating across all settings. DenseNet-121 and EfficientNet-B0 generally achieved the strongest results in MC classification for DTI and MRI, while B performance was often higher and more stable, particularly for MRI axial slices. PET exhibited the largest variability and the highest absolute performance, especially in coronal views, where DenseNet-121 reached an F1 (B) of 81.59 ± 3.40. Full results of the convolutional architectures are reported in Table 3. Table 4 systematically summarizes performance across all single-modality and single-slice experiments. Within DTI and MRI modalities, performance was largely comparable across both DinoV3 and RadioDino variants, with the top MC F1 for DTI axial achieved by DinoV3 small (39.76 ± 6.79) and the top B F1 for MRI axial achieved by DinoV3 base (70.92 ± 7.38). Nevertheless, individual model superiority varied by both modality and anatomical plane, and no approach dominated for every scenario highlighting the importance of multi-faceted evaluation.

PET, on the other hand, consistently revealed the largest performance gaps. For axial PET, RadioDino small achieved an F1 (B) of 78.90 ± 5.38, significantly outperforming both sizes of DinoV3 (best: 69.91 ± 5.51), while also posting competitive MC F1 (42.09 ± 7.39 in coronal with RadioDino base). This approximately 10% absolute F1 advantage in B PET tasks was especially pronounced across all anatomical axes.

5.4. 3D baseline comparison

To contextualize the proposed slice-based approach, we additionally compare our results against representative 3D baselines trained directly on full PET volumes, since these models achieve the strongest performance among unimodal configurations. In this analysis, all models were evaluated under the same data split and evaluation protocol. As reported in Table 6, conventional 3D architectures such as ResNet18 and ResNet50 exhibit competitive performance in the BCN vs. AD setting, with the best result achieved by a 3D ResNet50 (86.88 ± 2.49% F1), while the best multiclass result is achieved by 3D ResNet18 (55.04 ± 3.70% F1).

In addition to CNN-based comparators, we also evaluated Swin and ViT 3D baselines as representative transformer-based volumetric references. In our experiments, these transformer models proved substantially more sensitive to optimization choices than the 3D CNN baselines. In particular, without a dedicated transformer-oriented optimization setup, their performance collapsed, whereas under the final configuration reported in Table 6, Swin achieved 44.33 ± 1.71% F1 in the multiclass setting and 82.32 ± 4.21% in the binary setting, while ViT achieved 41.13 ± 2.91% and 83.29 ± 5.23%, respectively.

These results indicate that when full PET volumes are available and the task is restricted to a single modality, dedicated 3D models remain stronger than the present 2D formulation in this specific setting. Our slice-based design should therefore not be interpreted as a replacement for volumetric models. Instead, it was adopted primarily to preserve compatibility with currently available pretrained 2D foundation encoders and to maintain a unified experimental framework across modalities, fusion strategies, and missing-modality settings. This choice should be understood as a methodological design decision and, at the same time, as a limitation, since it discards full volumetric context. Differences in performance across orientations may partly reflect the anatomical structures captured in the selected central slices. The proposed framework is therefore intended primarily as a foundation-model-based benchmark for multimodal analysis under incomplete modality availability within a 2D input formulation.

These results should instead be interpreted in the context of the modeling objective of this work. The proposed framework was designed around pretrained 2D foundation encoders and a unified multimodal pipeline, rather than around maximizing PET-only volumetric performance. Consequently, the comparison with 3D baselines should be read as a boundary condition of the present design: when full PET volumes are available and the task is unimodal, dedicated 3D models remain a stronger choice. 3D ResNet architectures operate on full $96 \times 96 \times 96$ volumes, which is the default volume size for the MONAI-framework selected pretrained models.

Taken together, these observations suggest that while 3D models may achieve higher peak performance in the PET-only setting, the main advantage of the proposed framework lies in flexible multimodal analysis under incomplete modality availability rather than in outperforming the strongest unimodal volumetric baselines. Table 6 shows that dedicated 3D models remain stronger in the PET-only B setting, where full volumetric information is available, and the task is restricted to a single modality. The proposed framework is therefore not intended to replace these best-case unimodal baselines, but rather to address multimodal analysis under incomplete modality availability.

5.5. Multimodal results

As described in Table 1, we designed our experiments around four main scenarios: (i) using all slices from a single modality, (ii) fixing a slice type and using all modalities, (iii) using all slices from two modalities simultaneously, and (iv) combining all modalities and all slices together.

To improve training quality, we only included patients with at least one available modality. For missing data, we employed a learnable token serving as a [CLS] token. This strategy not only avoids limiting the

Table 5
Classification results for multi-modal analysis using CLS mean. Results are reported as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	F1 (MC) (%)	F1 (B) (%)
MRI-Full	RadioDino base	47.06 ± 3.91	72.41 ± 6.09
MRI-Full	RadioDino small	45.69 ± 4.49	72.94 ± 6.33
DTI-Full	RadioDino base	44.09 ± 4.13	58.02 ± 7.99
DTI-Full	RadioDino small	41.15 ± 4.68	61.95 ± 8.03
PET-Full	RadioDino base	**53.38 ± 4.64**	**82.48 ± 4.52**
PET-Full	RadioDino small	49.41 ± 5.86	81.92 ± 3.84
Axial-Cross	RadioDino base	53.33 ± 3.14	73.69 ± 4.19
Axial-Cross	RadioDino small	**54.54 ± 2.30**	**77.53 ± 4.32**
Sagittal-Cross	RadioDino base	50.50 ± 4.14	68.98 ± 5.05
Sagittal-Cross	RadioDino small	52.94 ± 4.14	71.91 ± 4.45
Coronal-Cross	RadioDino base	54.10 ± 3.23	72.35 ± 3.89
Coronal-Cross	RadioDino small	52.81 ± 3.81	76.74 ± 3.22
MRI-DTI	RadioDino base	43.81 ± 4.87	65.44 ± 3.79
MRI-DTI	RadioDino small	49.23 ± 3.70	65.78 ± 4.06
MRI-PET	RadioDino base	43.90 ± 4.65	**76.35 ± 4.65**
MRI-PET	RadioDino small	44.16 ± 4.16	74.23 ± 6.16
DTI-PET	RadioDino base	43.38 ± 4.91	63.48 ± 3.51
DTI-PET	RadioDino small	**52.97 ± 5.25**	69.91 ± 5.01
Full	RadioDino base	45.74 ± 3.53	71.32 ± 3.53
Full	RadioDino small	**49.42 ± 3.53**	**73.83 ± 3.53**

Table 6
Performance of 3D CNN and transformer-based models trained on full PET volumes. The updated Swin and ViT results show that volumetric transformer baselines can achieve competitive performance when properly optimized, although 3D ResNet models remain the strongest baselines overall in this PET-only setting. Results are reported for MC and B classification tasks as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	F1 (MC) (%)	F1 (B) (%)
PET-3D	DenseNet121	28.98 ± 3.26	62.77 ± 4.35
PET-3D	ResNet18	**55.04 ± 3.70**	85.21 ± 3.00
PET-3D	ResNet50	45.35 ± 5.36	**86.88 ± 2.49**
PET-3D	Swin	44.33 ± 1.71	82.32 ± 4.21
PET-3D	ViT	41.13 ± 2.91	83.29 ± 5.23

analysis to valid pairs but also provides a partial solution for handling incomplete data.

This token is prepended to the input sequence in place of the absent modality's CLS embedding, allowing the model to learn an appropriate representation for missing data. While this design provides a flexible mechanism for handling incomplete multimodal inputs, we acknowledge that the specific initialization and training dynamics of this token may influence performance. This mechanism was introduced as a practical missing-input handling strategy within the same fusion architecture and should not be interpreted as a dedicated latent-space aggregation baseline.

The first set of experiments was conducted using LoRA, where images were fed sequentially into RadioDino. For each image, the [CLS] token was extracted, and the average across all [CLS] tokens was computed to obtain the final prediction for each patient. This representation was then passed through a linear classification head. The complete results using mean averaging are reported in Table 5. Several patterns emerge: PET-based experiments consistently achieve higher F1 scores across different settings. Interestingly, the best results in the CN vs. MCI vs. AD task are obtained by using all axial slices across modalities, achieving 54.54%, which is only 1.16% higher than using all PET slices alone. These results suggest that PET may dominate the multimodal signal in the present cohort and input setting. This phenomenon is evident not only in multimodal experiments but also in the single-slice results reported in Section 5.3.

For the B classification task, PET again dominates, achieving the highest F1 score of 82.48% when using all PET slices. In contrast, adding further modalities does not consistently improve performance and can reduce it. For example, the Full multimodal setting with the Transformer head reaches 78.38%, remaining below PET-Full, while other multimodal combinations also show that performance depends strongly on the selected modality set. These results indicate that, in the present cohort and input setting, PET provides the strongest discriminative signal, whereas MRI and DTI do not always contribute sufficiently complementary information to offset the additional fusion complexity. A plausible explanation is that less informative or more incomplete modalities, together with the missing-modality token mechanism, introduce variability that the fusion head cannot fully suppress. Therefore, the proposed multimodal strategy should not be interpreted as a uniformly superior replacement for the strongest unimodal PET model. Instead, the contribution of the present framework is to provide a unified testbed for multimodal analysis under heterogeneous modality availability, where flexibility, controlled comparison across configurations, and explicit characterization of failure modes are central objectives alongside predictive performance.

Notably, multimodal results improve upon the single-slice baseline by 2.53 percentage points in the MC task and by 0.75 points in the B task. While the latter improvement may seem minor, in a screening context the proposed methodology may still support more flexible diagnostic modeling under incomplete-modality conditions.

However, when comparing across all scenarios, we observe that using all modalities with all slices together does not necessarily yield the best results. In several cases, this configuration underperforms compared to more targeted setups, suggesting that simply aggregating all available information may introduce noise or redundancy rather than improving classification performance.

To further investigate fusion strategies, we experimented with a Transformer-based head as described in Section 4.6. Given the high computational cost of training such a model across all possible configurations, we restricted this analysis to the best-performing CLS Mean configurations for both MC and B classification tasks, as summarized in Table 7. This choice allows the Transformer head experiments to serve as a practical and computationally tractable baseline for future extensions.

Overall, the Transformer fusion head demonstrates complementary behavior compared to the CLS Mean approach. While the CLS Mean head performs more robustly in simpler settings (e.g., single-modality

Table 7
Comparison of classification results for best-performing multimodal configurations, contrasting the CLS mean fusion head with the Transformer fusion head. The Transformer head shows a notable improvement in the most complex, fully multimodal scenarios, while the CLS Mean head performs better in single-modality and cross-slice fusion. Results are reported as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	Fusion Head	F1 (MC) (%)	F1 (B) (%)
PET-Full	RadioDino base	CLS Mean	**53.38 ± 4.64**	**82.48 ± 4.52**
		Transformer	44.34 ± 9.23	76.41 ± 5.00
Axial-Cross	RadioDino small	CLS Mean	**54.54 ± 2.30**	**77.53 ± 4.32**
		Transformer	51.15 ± 3.39	77.08 ± 3.63
DTI-PET	RadioDino small	CLS Mean	52.97 ± 5.25	72.80 ± 4.70
		Transformer	**54.89 ± 4.25**	74.10 ± 4.50
MRI-PET	RadioDino base	CLS Mean	48.20 ± 4.20	**76.35 ± 4.65**
		Transformer	50.10 ± 3.85	71.35 ± 3.35
Full	RadioDino small	CLS Mean	49.42 ± 3.53	73.83 ± 3.53
		Transformer	**54.20 ± 2.86**	**78.38 ± 4.44**

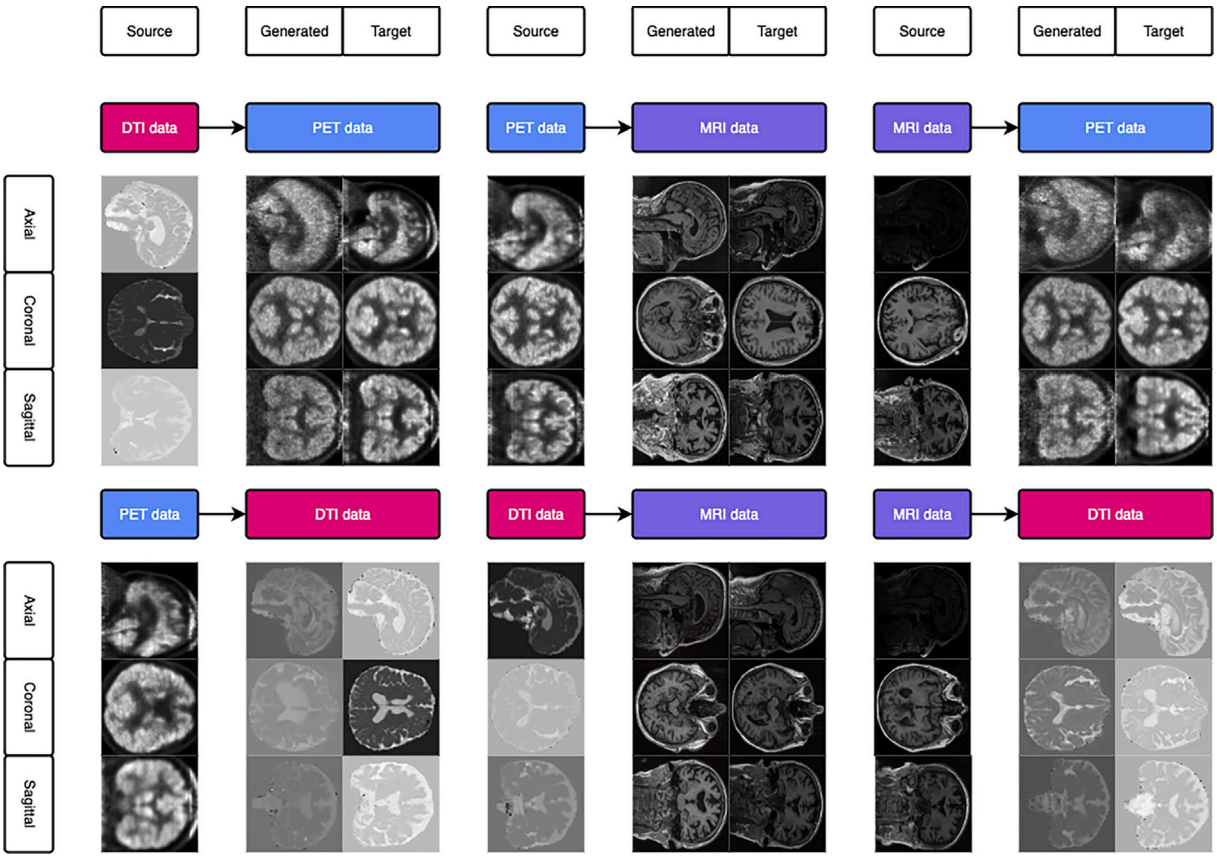


Fig. 4. Visual examples of modality-to-modality translation for generative data augmentation. Each column shows a source modality, the corresponding generated target, and the real target image. Rows represent different anatomical planes (axial, coronal, sagittal). The translation is performed using ControlNet-based conditioning between DTI, PET, and MRI data pairs. The generated images preserve anatomical structure while capturing modality-specific intensity patterns, demonstrating the model's ability to synthesize auxiliary signals, but its effect is strongly configuration-dependent, across modalities to support learning under incomplete multimodal data.

or cross-slice experiments), the Transformer head shows a consistent advantage in more complex multimodal scenarios. For instance, in the Full multimodal configuration, where all modalities and slice types are combined, the Transformer head outperforms the CLS Mean head by +4.78% in MC and +4.55% in the B classification tasks, achieving 54.20% and 78.38%, respectively. Despite this improvement over the corresponding Full CLS Mean configuration, the Full multimodal Transformer result in the B task remains below the strongest unimodal PET-Full baseline (82.48%), indicating that multimodal fusion is not uniformly beneficial in this setting. Both improvements are statistically significant according to a paired t-test when using all 25 runs, including the k-fold sub-runs (p_a: MC 1.60×10^{-6}; B 2.22×10^{-4}). When considering only the 5 independent repetitions, the MC improvement remains significant ($p_i = 0.047$), whereas the B improvement is no longer statistically significant ($p_i = 0.11$). Similarly, in the DTI–PET configuration, the Transformer fusion head achieves the best MC result overall (54.89%), suggesting that the model benefits from joint feature refinement when integrating complementary modalities.

5.6. Generative augmented classification

To provide a clear and effective mitigation for the missing data issue, we follow the reasoning that a patient must be associated with at least one imaging modality. Therefore, we train an imaging translation model to translate from one modality to another. By leveraging existing paired modalities, we use one as input for the ControlNet guide and the other as the ground truth for the generated modality. We iterate the training-generation process for each possible pair. The complete generation metrics on a held-out test set comprising 20% of the total data for each pair are reported in Table 11, along with several metrics commonly used in image generation evaluation. A visualization of the generated images on the test set is shown in Fig. 4.

The training hyperparameters are as follows: 20,000 iterations with a batch size of 32 and an image resolution of 224×224. We use the L2 loss as our guiding objective for the generative approach. Images were preprocessed via mean–standard deviation normalization using the ImageNet standard values and cropped such that the head region touches all borders to overcome pose differences among patients.

We use Table 11 as a guideline for selecting the best combination of generative models to fill missing data, guided by the average of all selected metrics. For metrics such as MSE, where lower values indicate better performance, we subtracted the metric value from 1, while for PSNR we applied a min–max normalization to scale values to the range $[0, 1]$.

We repeated the experiments reported in Table 7, this time including the generated data in the training set (Table 9). We observed no significant improvement in most metrics, except for the PET-Full experiments using the transformer-based head, which achieved a new highest F1 score of 56.41% in MC classification. When compared against the corresponding non-augmented PET-Full configuration using the transformer-based head, the observed improvement is statistically significant both across all 25 runs including k-fold sub-runs ($p_a = 9.23 \times 10^{-7}$) and across the 5 independent repetitions ($p_i = 0.04$) (Table. 8).

A clear observation is that a simple mean of the CLS tokens is not sufficient: while filling missing data provided more information, it also introduced noise, leading to poorer generalization capabilities. The transformer-based head, despite slightly lower performance, achieved results close to those in Table 7. This can be explained by data quantity: the PET modality has the largest number of samples among all modalities, so doubling its training data was beneficial. Conversely, for modalities such as MRI, where synthetic data exceeded real data, the scarcity of real samples led to poorer results. This degradation is substantial in several settings. In particular, the CLS Mean binary F1 dropped from 82.48 ± 4.52 to 65.54 ± 4.40 in PET-Full and from 77.53 ± 4.32 to 56.51 ± 5.06 in Axial-Cross, confirming that non-selective synthetic augmentation can be harmful when synthetic features compete with real ones.

5.6.1. Selective synthetic augmentation ablation

To better understand whether the observed degradation was caused by the indiscriminate use of synthetic modalities, we performed an additional ablation focused on the PET-Full configuration with the Transformer fusion head in the binary CN vs. AD setting. In this analysis, we compared synthetic-loss weighting through λ_{syn} and two usage policies: a non-selective policy, in which synthetic modalities were used even when the corresponding real modality was already available, and a missing-only policy, in which synthetic modalities were used only when the real modality was unavailable. Looking only at binary F1, the main pattern is that controlling when synthetic data are introduced is more important than merely reducing their loss contribution after inclusion. Under non-selective synthetic augmentation with loss weighting, performance increased from 70.61 ± 5.76 at $\lambda_{\text{syn}} = 0.25$ to 71.83 ± 5.23 at $\lambda_{\text{syn}} = 0.5$. When weighting was combined with missing-only usage, performance improved from 70.14 ± 4.07 at $\lambda_{\text{syn}} = 0.25$ to 72.17 ± 5.59 at $\lambda_{\text{syn}} = 0.5$. Thus, among the weighted variants, restricting synthetic modalities to genuinely missing cases was more effective than applying loss downweighting under non-selective augmentation. For reference, the corresponding PET-Full Transformer result reported in Table 9 is 75.86 ± 3.48, while the real-only Transformer baseline is 76.41 ± 5.00. Taken together, these results indicate that, in the PET-Full configuration with the Transformer fusion head, the main source of degradation is not simply the optimization weight assigned to synthetic samples, but their use when real modalities are already available. Selective use mitigates the degradation substantially, although it does not fully recover the real-only baseline.

Table 8
Ablation study on synthetic-data usage policy in the PET-Full configuration with the Transformer fusion head for binary classification. *Non-selective* denotes the use of synthetic modalities regardless of whether the corresponding real modality is already available, whereas *missing-only* restricts synthetic modalities to cases where the real modality is unavailable. The results show that controlling when synthetic data are introduced is more effective than loss downweighting alone, with missing-only usage consistently outperforming the corresponding non-selective weighted setting.

Synthetic-data policy	λ_{syn}	Test F1 (B) (%)
Non-selective	0.25	70.61 ± 5.76
Non-selective	0.50	71.83 ± 5.23
Missing-only	0.25	70.14 ± 4.07
Missing-only	0.50	**72.17 ± 5.59**

Table 9
Comparison of classification results using the mean and transformer fusion heads after augmenting the training data with synthetically generated modalities. The transformer head consistently outperforms the Mean head across all configurations, particularly in PET-based and fully multimodal setups. These results show that the effect of generative augmentation depends strongly on the fusion strategy and configuration, with clear degradation for CLS Mean and more stable behavior for the Transformer head. Results are reported as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	Fusion Head	F1 (MC) (%)	F1 (B) (%)
PET-Full	RadioDino base	CLS Mean	42.57 ± 4.62	65.54 ± 4.40
		Transformer	**56.41 ± 3.55**	**75.86 ± 3.48**
Axial-Cross	RadioDino small	CLS Mean	36.45 ± 2.22	56.51 ± 5.06
		Transformer	**49.40 ± 3.02**	**72.89 ± 3.37**
DTI-PET	RadioDino small	CLS Mean	39.18 ± 3.25	60.80 ± 4.10
		Transformer	**54.35 ± 3.40**	63.95 ± 3.90
MRI-PET	RadioDino base	CLS Mean	40.22 ± 3.50	62.54 ± 3.36
		Transformer	44.10 ± 3.20	**78.09 ± 3.36**
Full	RadioDino small	CLS Mean	39.33 ± 2.56	64.64 ± 4.24
		Transformer	**53.49 ± 3.66**	**74.96 ± 3.48**

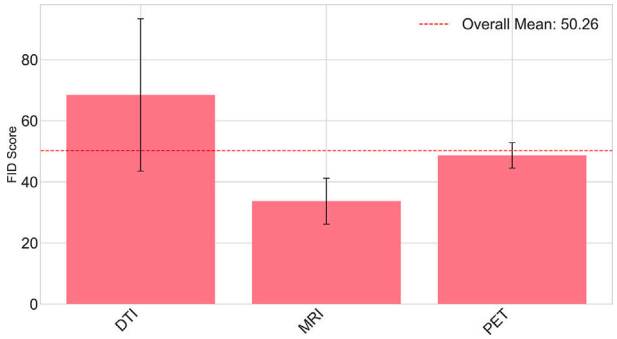


Fig. 5. FID analysis across imaging modalities and slice orientations. The left panel reports mean FID scores aggregated by modality, while the right panel shows aggregation by orientation. Error bars indicate variability across orientations or modalities, respectively. The dashed red line denotes the global mean FID. PET exhibits relatively stable FID values, whereas MRI achieves a lower average FID but with increased sensitivity to generative augmentation. DTI shows higher variability, particularly across orientations, highlighting orientation-dependent distribution mismatch in the generated data.

The FID analysis in Fig. 5 strongly supports this interpretation. PET maintains consistent FID scores across orientations (axial: 46.29, sagittal: 46.28, coronal: 53.49), indicating that the generative process preserves the underlying data distribution and limits the introduction of noise. In contrast, MRI achieves lower absolute FID values (axial: 33.20, sagittal: 41.37, coronal: 26.42) yet fails to improve generalization, demonstrating that visual fidelity alone is insufficient and may mask distributional bias. DTI further highlights this effect through pronounced orientation-dependent FID degradation, particularly in the sagittal view (axial: 44.69, sagittal: 94.35, coronal: 66.25). These results confirm that effective generative augmentation depends on distributional consistency rather than absolute FID minimization.

The mixed downstream results indicate that synthetic augmentation can sometimes provide a useful auxiliary signal, but its effect is strongly configuration-dependent. In the classification experiments, test sets consisted solely of real images, while synthetic data were used during training and validation. The subject-level split was defined before any generative training. Accordingly, for each outer fold, a separate ControlNet model was trained exclusively on subjects belonging to that fold's training partition and was used only to synthesize missing modalities for training and validation subjects within the same fold. Subjects assigned to the corresponding classification test fold were never seen during generative training or synthetic data generation. Given computational constraints, the generative models were trained once per fold and

Table 10
Performance of k-nearest neighbors imputation for estimating missing modalities across different experimental settings. Results compare CLS mean and transformer-based fusion heads.

Experiment	Model	Fusion Head	F1 (MC) (%)	F1 (B) (%)
PET-Full	RadioDino base	CLS Mean	**45.12 ± 3.95**	**70.05 ± 4.05**
		Transformer	42.30 ± 5.10	67.85 ± 4.90
Axial-Cross	RadioDino small	CLS Mean	**38.12 ± 2.00**	**69.85 ± 4.80**
		Transformer	36.45 ± 2.85	68.20 ± 4.15
DTI-PET	RadioDino small	CLS Mean	40.65 ± 3.05	65.50 ± 3.95
		Transformer	**42.10 ± 3.80**	**67.25 ± 4.10**
MRI-PET	RadioDino base	CLS Mean	**41.55 ± 3.25**	**68.45 ± 3.20**
		Transformer	40.20 ± 4.10	66.30 ± 3.75
Full	RadioDino small	CLS Mean	41.70 ± 2.45	69.85 ± 4.05
		Transformer	**43.95 ± 3.10**	**71.90 ± 4.40**

reused across repetitions rather than training separate models for each iteration or subfold. Therefore, we trained five generative models for each experimental setup, corresponding to the five cross-validation folds. The observed improvements in classification performance demonstrate that even imperfectly reconstructed images can provide valuable information for the learning process, while evaluation on real images ensures reliable assessment of model performance. As an additional downstream validation, we trained single-modality classifiers using only synthetically generated images and evaluated them on held-out real images from the corresponding target modality. Performance was consistently lower than the corresponding real-data baselines reported in Table 4. In the binary B setting, the best synthetic-only F1 scores were 50.84 ± 9.67 for DTI, 56.29±8.34 for MRI, and 39.27±5.33 for PET. In the multiclass MC setting, the best synthetic-only F1 scores were 24.18±5.90 for DTI, 34.26±7.07 for MRI, and 19.96 ± 5.67 for PET. These results indicate that the generated images retain limited disease-related signal, but they are not sufficient as stand-alone substitutes for real acquisitions, with the strongest degradation observed for PET and for the multiclass setting. Accordingly, in the present framework synthetic modalities should be interpreted only as auxiliary training-time augmentation for improving robustness under missing-modality settings, rather than as independently reliable data for direct model training or clinical interpretation. The role of the generative model within our framework should therefore be understood as a limited training-time augmentation strategy for missing-modality settings. The additional synthetic-only downstream validation shows that classifiers trained exclusively on generated images underperform their real-data counterparts across all modalities and tasks, indicating that the synthetic images are not suitable as stand-alone replacements for real acquisitions or for direct clinical interpretation.

We evaluated whether the generative model preserves disease-relevant patterns by stratifying SSIM scores by diagnosis (CN, MCI, AD). Table 12 reports the results for each source–target pair and orientation. While absolute SSIM values are modest for some cross-modality translations, the scores are broadly comparable across diagnostic groups, suggesting that the model does not introduce strong bias or systematically degrade disease-related information.

To contextualize the impact of generative augmentation, we additionally compare ControlNet-based synthesis against simpler missing-modality imputation strategies commonly used as baselines under the same training hyperparameters, specifically k-NN. In k-NN imputation, missing modality features are estimated by averaging the available representations of the most similar patients, with similarity computed using the embeddings from an Inception network, consistent with Fréchet Inception Distance evaluations; specifically, the similarity is averaged across available modalities. As reported in Table 10, these approaches provide modest or even lower performance compared to the no-imputation baseline Table 7, remaining consistently inferior to generative augmentation. This indicates that the performance gains from ControlNet are not due merely to feature averaging or local similarity-based imputation.

5.7. Generative augmented classification leveraging clinical data

Building upon the generative augmentation experiments, we extended our framework by integrating clinical and demographic variables into the multimodal fusion process. The inclusion of tabular patient information, such as demographic and cognitive assessment data, provides complementary cues that enhance disease characterization beyond imaging-based biomarkers. This step aims to capture patient-level variability and improve the robustness of Alzheimer's disease classification.

The results obtained without synthetic augmentation are summarized in Table 13. Consistent with previous multimodal analyses, the Transformer-based fusion head continues to outperform the CLS Mean head across nearly all configurations. For instance, in the *Full* multimodal setting, the Transformer head achieved 52.30% (MC) and 79.66% (B), surpassing the CLS Mean head by +7.36 and +15.92 percentage points, respectively. Similar trends were observed across modality-pair and cross-slice configurations, confirming that the attention-driven fusion mechanism provides more expressive cross-modal reasoning when

Table 11
Modality translation metrics Mean ± Std computed across test set images.

Source	Target	Orientation	SSIM ↑	PSNR ↑	MSE ↓	RMSE ↓	MAE ↓	NCC ↑
DTI	MRI	Axial	0.18 ± 0.06	12.92 ± 2.15	0.06 ± 0.03	0.23 ± 0.06	0.18 ± 0.06	0.34 ± 0.10
DTI	MRI	Coronal	0.22 ± 0.06	12.45 ± 1.74	0.06 ± 0.03	0.24 ± 0.05	0.18 ± 0.05	0.44 ± 0.11
DTI	MRI	Sagittal	0.15 ± 0.07	13.02 ± 2.39	0.06 ± 0.03	0.23 ± 0.07	0.17 ± 0.06	0.25 ± 0.11
DTI	PET	Axial	0.20 ± 0.04	11.99 ± 0.87	0.06 ± 0.01	0.25 ± 0.03	0.20 ± 0.03	0.42 ± 0.09
DTI	PET	Coronal	0.25 ± 0.04	13.63 ± 0.69	0.04 ± 0.01	0.21 ± 0.02	0.16 ± 0.02	0.66 ± 0.05
DTI	PET	Sagittal	0.22 ± 0.04	12.27 ± 0.98	0.06 ± 0.02	0.25 ± 0.03	0.19 ± 0.03	0.47 ± 0.09
MRI	DTI	Axial	0.45 ± 0.24	13.09 ± 5.27	0.09 ± 0.11	0.26 ± 0.16	0.25 ± 0.22	0.37 ± 0.59
MRI	DTI	Coronal	0.44 ± 0.28	12.59 ± 7.20	0.14 ± 0.15	0.31 ± 0.21	0.30 ± 0.24	0.32 ± 0.39
MRI	DTI	Sagittal	0.38 ± 0.27	11.27 ± 6.09	0.15 ± 0.14	0.34 ± 0.19	0.33 ± 0.20	0.34 ± 0.28
MRI	PET	Axial	0.21 ± 0.04	12.31 ± 0.89	0.06 ± 0.01	0.24 ± 0.03	0.19 ± 0.03	0.47 ± 0.08
MRI	PET	Coronal	0.26 ± 0.04	13.56 ± 0.81	0.04 ± 0.01	0.21 ± 0.02	0.16 ± 0.02	0.67 ± 0.05
MRI	PET	Sagittal	0.22 ± 0.05	12.45 ± 0.84	0.06 ± 0.01	0.24 ± 0.02	0.19 ± 0.02	0.49 ± 0.09
PET	DTI	Axial	0.47 ± 0.22	13.93 ± 5.49	0.08 ± 0.10	0.24 ± 0.15	0.23 ± 0.21	0.43 ± 0.85
PET	DTI	Coronal	0.53 ± 0.22	14.46 ± 5.61	0.07 ± 0.09	0.23 ± 0.14	0.22 ± 0.15	0.39 ± 0.16
PET	DTI	Sagittal	0.39 ± 0.28	12.33 ± 5.99	0.12 ± 0.13	0.30 ± 0.18	0.29 ± 0.26	0.37 ± 1.21
PET	MRI	Axial	0.17 ± 0.05	13.13 ± 1.67	0.05 ± 0.02	0.22 ± 0.04	0.16 ± 0.04	0.35 ± 0.10
PET	MRI	Coronal	0.20 ± 0.07	13.34 ± 1.76	0.05 ± 0.03	0.22 ± 0.05	0.16 ± 0.05	0.44 ± 0.09
PET	MRI	Sagittal	0.13 ± 0.05	13.40 ± 2.11	0.05 ± 0.03	0.22 ± 0.06	0.17 ± 0.06	0.25 ± 0.10

Table 12
Modality translation metrics: Mean SSIM ± Std computed across test set images stratified by diagnosis.

Source	Target	Orientation	CN	MCI	AD
DTI	MRI	Axial	0.18 ± 0.05	0.18 ± 0.05	0.19 ± 0.05
DTI	MRI	Coronal	0.22 ± 0.06	0.22 ± 0.04	0.22 ± 0.05
DTI	MRI	Sagittal	0.14 ± 0.06	0.14 ± 0.06	0.18 ± 0.08
DTI	PET	Axial	0.20 ± 0.04	0.20 ± 0.04	0.20 ± 0.03
DTI	PET	Coronal	0.26 ± 0.04	0.25 ± 0.04	0.25 ± 0.04
DTI	PET	Sagittal	0.23 ± 0.04	0.22 ± 0.04	0.23 ± 0.04
MRI	DTI	Axial	0.45 ± 0.24	0.39 ± 0.28	0.53 ± 0.20
MRI	DTI	Coronal	0.40 ± 0.28	0.42 ± 0.30	0.46 ± 0.26
MRI	DTI	Sagittal	0.40 ± 0.27	0.39 ± 0.27	0.43 ± 0.29
MRI	PET	Axial	0.22 ± 0.04	0.22 ± 0.04	0.20 ± 0.04
MRI	PET	Coronal	0.26 ± 0.05	0.26 ± 0.04	0.27 ± 0.04
MRI	PET	Sagittal	0.23 ± 0.05	0.22 ± 0.04	0.22 ± 0.04
PET	DTI	Axial	0.48 ± 0.21	0.46 ± 0.23	0.43 ± 0.28
PET	DTI	Coronal	0.55 ± 0.24	0.56 ± 0.20	0.58 ± 0.18
PET	DTI	Sagittal	0.42 ± 0.27	0.38 ± 0.29	0.38 ± 0.30
PET	MRI	Axial	0.17 ± 0.04	0.17 ± 0.05	0.18 ± 0.05
PET	MRI	Coronal	0.19 ± 0.05	0.20 ± 0.07	0.20 ± 0.08
PET	MRI	Sagittal	0.13 ± 0.04	0.13 ± 0.05	0.12 ± 0.05

combining imaging and tabular data. However, statistical evaluation using both the full set of 25 runs and the 5 independent repetitions indicates that these differences are not statistically significant ($p > 0.05$) for either MC or B classification. Therefore, while clinical features may provide additional context, their contribution to overall performance is limited in the present experiments.

Table 14 reports the corresponding results obtained when synthetic data generated through ControlNet was incorporated into the training set. Based on the consistent superiority of the Transformer head in earlier experiments, we restricted this analysis to the Transformer configuration only. This choice ensures a computationally efficient setup while maintaining methodological coherence with prior findings.

Interestingly, the addition of synthetic modalities led to moderate improvements in most configurations, particularly in PET-Full (57.80% MC, 81.40% B) and MRI-PET (80.49% B). These results demonstrate that generative augmentation can further enhance multimodal fusion when guided by reliable clinical embeddings. A separate observation concerns the multiclass setting. As shown in Table 14, the proposed framework remains competitive in MC classification and, in the best multimodal configuration, attains higher MC F1 than the 3D PET baselines. This suggests that its main advantage emerges in more complex multimodal classification settings rather than in the best-case unimodal binary scenario.

Taken together, Tables 9 and 14 delineate the operating boundary of generative augmentation in the present framework. The most favorable regime is PET-centered incomplete-data settings, where a strong real PET signal remains available and the synthetic samples act as a limited complement rather than the dominant source of information. In these cases, augmentation improves PET-Full in the multiclass setting and, when clinical variables are included, also benefits PET-Full and MRI-PET in the binary setting. By contrast, performance becomes more mixed as missingness handling requires broader multimodal completion: DTI-PET shows task-dependent behavior, and the fully multimodal Full configuration does not show a uniform advantage, with MC performance decreasing after augmentation. Most notably, the Axial-Cross configuration degrades with augmentation in both Sections 5.6 and Section 5.7, indicating that this setting should not be enabled by default in the present implementation. Overall, these findings suggest that the proposed framework is most appropriately interpreted as a unified benchmark for studying multimodal learning under incomplete-modality conditions, rather than as a uniformly superior alternative to strong unimodal baselines.

Table 13
Comparison of multimodal classification performance with integrated clinical features using both Mean and Transformer fusion heads. The Transformer head consistently outperforms the CLS Mean approach across all configurations, demonstrating its superior ability to capture cross-modal dependencies between imaging and clinical data. Results are reported as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	Fusion Head	F1 (MC) (%)	F1 (B) (%)
PET-Full	RadioDino base	CLS Mean	37.12 ± 4.50	67.44 ± 4.37
		Transformer	51.94 ± 4.89	76.41 ± 4.74
Axial-Cross	RadioDino small	CLS Mean	44.50 ± 4.12	64.25 ± 2.95
		Transformer	52.59 ± 3.04	77.83 ± 4.15
DTI-PET	RadioDino small	CLS Mean	41.65 ± 3.10	61.40 ± 3.80
		Transformer	**54.25 ± 2.75**	63.10 ± 4.05
MRI-PET	RadioDino base	CLS Mean	39.52 ± 3.45	62.44 ± 4.47
		Transformer	42.80 ± 3.20	74.24 ± 3.62
Full	RadioDino small	CLS Mean	44.94 ± 3.05	63.74 ± 3.07
		Transformer	52.30 ± 2.80	**79.66 ± 4.75**

Table 14
Classification results using the Transformer fusion head after integrating both clinical features and synthetically generated imaging modalities. The results remain strongest in PET-centered configurations, but the effect of generative augmentation is mixed across settings and should not be interpreted as uniformly beneficial. Results are reported as mean ± standard deviation across repeated stratified group k-fold cross-validation with 5 folds and 5 repetitions, yielding 25 model fits per configuration.

Experiment	Model	F1 (MC) (%)	F1 (B) (%)
PET-Full	RadioDino base	**57.80 ± 4.65**	**81.40 ± 3.76**
Axial-Cross	RadioDino small	50.04 ± 2.58	71.71 ± 3.91
DTI-PET	RadioDino small	52.02 ± 3.61	74.85 ± 4.20
MRI-PET	RadioDino base	49.10 ± 3.55	80.49 ± 4.41
Full	RadioDino small	50.53 ± 2.75	80.17 ± 5.19

We also evaluated a gated variant of the Transformer fusion head in the most challenging setting, namely the Full configuration with clinical features and synthetic modalities, using RadioDino small under the same training protocol as the corresponding experiment reported in Table 14. In this variant, only the modality tokens were modulated before self-attention, while the CLS token, the clinical token, and the missing-modality masking mechanism were left unchanged. Let $X_{\text{mod}} = \{x_1, \ldots, x_M\}$ denote the set of modality tokens, with $x_i \in \mathbb{R}^{d_{\text{model}}}$. For each modality token, a scalar gate was computed as

$$g_i = \sigma\big(\mathbf{W}_2\, \phi\big(\mathbf{W}_1\, \mathrm{LN}(x_i) + \mathbf{b}_1\big) + \mathbf{b}_2\big),$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, $\phi(\cdot)$ is the GELU activation, and $\sigma(\cdot)$ is the sigmoid function. The gated modality token was then obtained as

$$\tilde{x}_i = g_i\, x_i,$$

and the final input sequence to the Transformer was formed by replacing each modality token x_i with its gated counterpart $\tilde{x}_i$. This experiment was restricted to the Full setting because it represents the most challenging fusion scenario, combining all imaging modalities, clinical variables, and synthetic augmentation, and therefore provides the most direct test of whether explicit token-wise suppression can mitigate noisy or weakly informative modalities. However, the gated variant did not improve over the standard Transformer, yielding 47.12 ± 1.93% in MC and 79.97 ± 4.27% in B classification, compared with 50.53 ± 2.75% and 80.17 ± 5.19% for the corresponding non-gated configuration. These results suggest that, in the present setting, this simple token-wise

Table 15
Demographic fairness analysis of RadioDino model configurations across neuroimaging modalities. Results show F1-scores, Calibration Gap (CG), and sample sizes (mean ± standard deviation) for classification stratified by ethnicity and gender. Sample sizes are averaged across the 5 iterations. Only subgroup-task combinations with at least 30 samples are reported quantitatively. Other ethnic subgroups did not satisfy this minimum sample-size criterion and are therefore omitted from the table.

Experiment	Model	Demographic Group	F1 (MC) (%)↑	CG (MC) (%)↓	N (MC)	F1 (B) (%)↑	CG (B) (%)↓	N (B)
PET-Full	RadioDino Base	White	**52.3 ± 5.1**	**8.4 ± 2.8**	135.8 ± 3.06	**77.4 ± 4.9**	**6.8 ± 2.6**	67.0 ± 1.29
		Male	**51.6 ± 5.8**	**9.0 ± 2.9**	81.4 ± 6.21	**76.8 ± 7.2**	8.1 ± 3.2	40.8 ± 2.08
		Female	51.5 ± 5.8	9.4 ± 3.6	65.6 ± 6.21	72.2 ± 8.3	**7.7 ± 3.5**	33.2 ± 2.08
Axial-Cross	RadioDino Small	White	52.1 ± 3.0	**8.5 ± 2.1**	188.4 ± 2.20	77.3 ± 5.1	**5.3 ± 2.7**	105.8 ± 2.53
		Male	49.4 ± 4.7	10.0 ± 3.5	113.0 ± 3.71	74.9 ± 4.0	5.8 ± 3.0	59.8 ± 4.76
		Female	**54.8 ± 6.0**	**8.4 ± 3.1**	101.0 ± 3.71	**78.5 ± 7.6**	**5.3 ± 2.4**	67.2 ± 4.76
DTI-PET	RadioDino Small	White	53.5 ± 2.4	**7.5 ± 2.4**	175.6 ± 3.74	79.0 ± 11.7	**21.0 ± 2.0**	93.22 ± 2.32
		Male	51.7 ± 4.4	8.6 ± 2.7	101.0 ± 7.10	74.6 ± 6.1	18.9 ± 4.8	101.0 ± 7.10
		Female	**54.8 ± 4.9**	**8.5 ± 3.5**	99.0 ± 7.10	**77.9 ± 6.7**	**17.2 ± 4.3**	99.0 ± 7.10
MRI-PET	RadioDino Base	White	52.0 ± 4.0	**5.6 ± 2.7**	77.2 ± 0.41	74.3 ± 4.0	**5.6 ± 2.7**	77.2 ± 0.41
		Male	49.5 ± 4.4	**5.8 ± 3.4**	45.6 ± 1.53	**75.5 ± 4.4**	**5.8 ± 3.4**	45.6 ± 1.53
		Female	**51.6 ± 7.7**	6.6 ± 3.3	39.4 ± 1.53	71.6 ± 7.7	6.6 ± 3.3	39.4 ± 1.53
Full	RadioDino Small	White	**51.9 ± 3.0**	**8.2 ± 2.6**	188.4 ± 2.20	79.3 ± 4.8	**6.4 ± 2.8**	105.8 ± 2.53
		Male	50.1 ± 5.2	9.0 ± 3.5	113.0 ± 3.71	77.9 ± 5.6	**6.8 ± 2.3**	59.8 ± 4.76
		Female	**53.2 ± 6.7**	**8.3 ± 2.4**	101.0 ± 3.71	**78.6 ± 8.4**	6.9 ± 3.3	67.2 ± 4.76

Table 16
Age-group performance analysis of RadioDino model configurations across neuroimaging modalities. Results show F1-scores, Calibration Gap (CG), and sample sizes (mean ± standard deviation) for classification. Only subgroup-task combinations with at least 30 samples are reported quantitatively. Other age bins did not satisfy this minimum sample-size criterion and are therefore discussed in the text rather than reported in the table. Cells marked with – correspond to subgroup-task combinations that did not satisfy the minimum sample-size criterion and were therefore not reported quantitatively.

Experiment	Model	Age	F1 (MC) (%)↑	CG (MC) (%)↓	N (MC)	F1 (B) (%)↑	CG (B) (%)↓	N (B)
PET-Full	RadioDino Base	80–89	50.3 ± 10.4	10.9 ± 4.6	58.8 ± 1.35	–	–	–
		90–99	45.3 ± 6.7	**8.6 ± 3.1**	47.2 ± 3.32	72.5 ± 8.8	8.8 ± 3.10	30.6 ± 4.77
Axial-Cross	RadioDino Small	70–79	55.8 ± 6.5	12.9 ± 5.5	50.8 ± 6.93	81.2 ± 9.6	4.6 ± 2.6	33.20 ± 6.14
		80–89	50.2 ± 4.9	10.4 ± 3.2	75.6 ± 3.45	80.4 ± 5.7	7.3 ± 3.2	40.40 ± 4.68
		90–99	41.4 ± 4.2	**9.9 ± 4.0**	64.8 ± 9.05	66.6 ± 10.7	7.1 ± 4.2	37.00 ± 5.52
DTI-PET	RadioDino Small	70–79	53.4 ± 8.1	9.7 ± 4.3	52.0 ± 5.74	–	–	–
		80–89	54.8 ± 3.3	**9.6 ± 3.0**	70.6 ± 4.73	72.8 ± 3.6	**6.7 ± 2.4**	35.30 ± 2.36
		90–99	40.4 ± 5.7	11.9 ± 4.3	55.2 ± 3.62	–	–	–
MRI-PET	RadioDino Base	90–99	74.0 ± 5.8	6.7 ± 2.8	38.6 ± 3.20	74.0 ± 5.8	6.7 ± 2.8	38.60 ± 3.20
Full	RadioDino Small	70–79	59.0 ± 7.0	11.2 ± 4.2	50.8 ± 6.93	82.7 ± 9.5	**4.6 ± 2.0**	33.20 ± 6.14
		80–89	47.9 ± 5.4	**9.6 ± 4.0**	75.6 ± 3.45	78.6 ± 6.7	6.8 ± 2.9	40.40 ± 4.68
		90–99	43.3 ± 5.3	10.4 ± 3.7	64.8 ± 9.05	71.8 ± 9.8	8.5 ± 4.2	37.00 ± 5.51

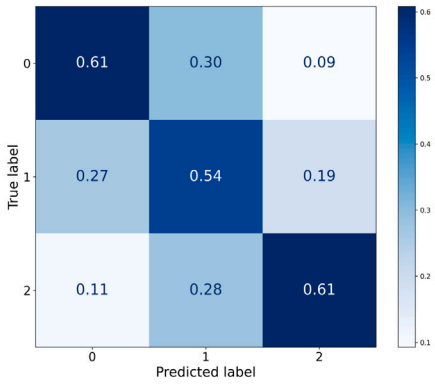
gating formulation is not sufficient to resolve the observed multimodal interference.

To provide deeper insight into model behavior beyond aggregate F1-scores, we report normalized confusion matrices for both the multiclass MC setting and the corresponding binary classification task Section 6. All confusion matrices are row-normalized, thus reflecting per-class recall and mitigating the effect of class imbalance. In the multiclass setting, the confusion matrix reveals the expected difficulty in discriminating adjacent disease stages. Correct classification rates are highest for CN and AD, while MCI exhibits increased confusion with both neighboring classes, consistent with its clinically heterogeneous and transitional nature. Misclassifications predominantly occur between CN and MCI, and between MCI and AD, whereas direct confusion between CN and AD remains limited. These patterns support the use of F1-score as a balanced metric while highlighting class-specific failure modes. The binary confusion matrix further confirms that performance gains are not driven by a single dominant class, but instead reflect balanced sensitivity across disease groupings.

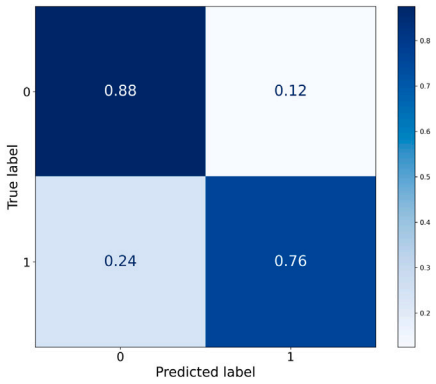
While these values are reported as a benchmark within our experimental framework, they should be interpreted in the context of the intrinsic difficulty of ADNI classification. Reported multiclass F1-scores in prior ADNI studies vary widely depending on cohort selection, modality availability, and evaluation protocol, and are often substantially lower than their binary counterparts. In this regard, the obtained result is consistent with the upper range of performance reported in approaches [5] under strict cross-validation and patient-level separation, rather than indicating a definitive state-of-the-art outcome.

6. Demographic and fairness analysis

For the assessment of fairness and generalizability, we adopted the multimodal configuration trained with synthetic data using the Transformer fusion head, selected for its consistent performance across previous experiments and its balance between representational richness and computational feasibility. This setup integrates both real and generated imaging modalities along with clinical features, enabling a comprehensive evaluation across demographic and diagnostic subgroups. To reduce the risk of over-interpreting statistically unstable subgroup estimates, the quantitative fairness analysis reported below is restricted to subgroup-task combinations with at least 30 samples. Subgroups not satisfying this criterion are omitted from the tables and are not used for quantitative fairness claims. To this end, we performed a subgroup analysis stratified by ethnicity, gender, and age, with Tables 15 and 16 reporting F1-scores and corresponding sample sizes across multimodal configurations. Specifically, ethnic and gender fairness is analyzed in Section 6.1, highlighting performance differences



(a) Normalized MC confusion matrix for the PET-Full MC approach reported in table 14.



(b) Normalized B confusion matrix for the PET-Full approach reported in table 14.

Fig. 6. (a) Normalized MC confusion matrix for the PET-Full MC approach reported in Table 14. (b) Normalized B confusion matrix for the PET-Full approach reported in Table 14.

across ethnicity and sex. Age-related trends are examined in Section 6.2, showing performance variations across age cohorts. Finally, overall interpretation and insights regarding demographic imbalances and their impact on model reliability are discussed in Section 6.4.

6.1. Ethnic and gender fairness

Across all experiments, the *White* subgroup exhibited the most stable and highest F1-scores, achieving 52.0% ± 5.0 in MC and 77.0% ± 5.0 in B classification for the *PET-Full* configuration. This result aligns with the substantially larger sample size available for this population ($N_{CN+MCI+AD}$ = 135.8 ± 3.06), suggesting that model performance strongly correlates with data availability.

Gender analysis revealed minimal bias between male and female subgroups, with comparable MC F1-scores across all modalities. For instance, in the *PET-Full* setting, both male and female participants achieved approximately 52.0% MC F1, while B classification slightly favored males (76.8% vs. 72.2%). The *Axial-Cross* and *Full* configurations showed marginally higher performance for females in the B setting, whereas the remaining differences were modest relative to the variability across configurations. Overall, among the subgroup comparisons supported by sufficient sample size, sex-related disparities appear limited.

6.2. Age-related trends

The age-group analysis reported in Table 16 is likewise restricted to subgroup-task combinations with at least 30 samples. Within the reported strata, performance is generally more stable in the 70–99 age range, whereas younger and extreme-age bins are not shown when the available sample size is insufficient for robust estimation. This pattern indicates that age-related differences should be interpreted jointly with the cohort distribution.

6.3. Algorithmic bias and potential clinical consequences

Beyond performance disparities, algorithmic bias may have tangible clinical consequences if deployed without appropriate safeguards. In particular, systematic underperformance for specific demographic or acquisition-defined subgroups could lead to delayed diagnosis, misclassification, or reduced access to appropriate follow-up care for underrepresented populations. In the context of neurodegenerative disease screening, such errors may exacerbate existing healthcare inequities by disproportionately affecting groups that are already less represented in large-scale datasets such as ADNI. While the present study focuses on identifying and quantifying subgroup-level performance differences, we emphasize that fairness-aware deployment requires additional mitigation strategies. Potential directions include the adoption of group-specific decision thresholds to balance sensitivity and specificity across subgroups, reweighting or oversampling underrepresented cohorts during training, and post hoc calibration to reduce systematic prediction gaps.

6.4. Interpretation

The fairness analysis highlights that model reliability is strongly influenced by demographic imbalances in the ADNI dataset. In particular, the limited representation of several ethnic categories and of some age extremes constrains the set of subgroup comparisons that can be supported quantitatively. These results underscore the need for more balanced recruitment strategies and for future multimodal foundation model studies explicitly designed around robust subgroup evaluation.

7. Limitations and future directions

While our framework provides a structured basis for studying multimodal AD diagnosis through foundation models and generative augmentation, several important considerations guide future research directions and clinical translation efforts. Dataset scope and generalizability are discussed in Section 7.1, emphasizing the need for more diverse populations. Architectural design choices are considered in Section 7.2, highlighting the trade-off between 2D slice-based processing and full 3D analysis. Synthetic data generation and biological fidelity are addressed in Section 7.3, identifying opportunities for more sophisticated generative models. Clinical data integration pathways are examined in Section 5.7, focusing on interpretability and practical deployment. Methodological focus and baseline strategy are detailed in Section 7.5, supporting targeted evaluation of core hypotheses. Fairness and representation are analyzed in Section 7.6, stressing bias-aware model development. Finally, translational opportunities are outlined in Section 7.7, providing a roadmap toward clinical adoption.

7.1. Dataset scope and generalizability

Our investigation leverages the ADNI dataset, recognized as the premier resource for AD neuroimaging research, which provides standardized protocols and longitudinal follow-up essential for robust model development. However, ADNI's carefully curated cohort, while methodologically advantageous, represents a demographically homogeneous population that may limit immediate generalization to global healthcare settings. This characteristic presents an opportunity to extend our framework to more diverse populations and validate its robustness across different demographic and clinical contexts. The controlled nature of

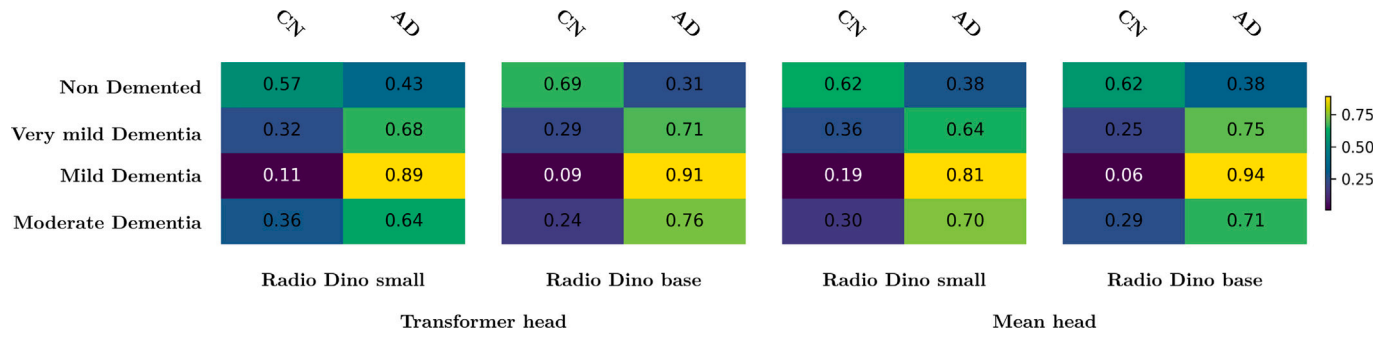


Fig. 7. Class-wise prediction distributions of the full MRI model evaluated on the OASIS-1 dataset, using both transformer and mean aggregation heads with RadioDino backbones.

ADNI data also motivates future work incorporating real-world clinical datasets with greater heterogeneity in imaging protocols, comorbidities, and disease presentations.

While ADNI provides a well-curated benchmark, it is characterized by limited demographic and site diversity, which constrains the assessment of model robustness across heterogeneous clinical settings. To preliminarily explore whether the observed trends extend beyond this homogeneous context, we evaluated the full MRI model on the independent OASIS-1 cohort [49], which differs from ADNI in terms of subject recruitment, acquisition protocols, and cohort composition. To reduce ambiguity while acknowledging the label mismatch, we additionally considered a simplified binary mapping for this auxiliary external analysis, in which OASIS-1 "Non Demented" subjects were mapped to CN and OASIS-1 "Very Mild Dementia", "Mild Dementia", and "Moderate Dementia" subjects were grouped into AD. Under this coarse mapping, the evaluated configurations yielded approximate macro-F1 values ranging from 0.581 to 0.669; the strongest configuration reached a macro-F1 of 0.669, an AD-class F1 of 0.552, a balanced accuracy of 0.729, and an overall accuracy of 0.710. These values were derived from aggregated confusion matrices and should therefore be interpreted only as approximate external diagnostic-consistency indicators rather than as directly comparable replications of the ADNI benchmark. As shown in Fig. 7, we also report class-wise prediction distributions, since the diagnostic label definitions in OASIS-1 still do not directly align with those used in ADNI, preventing a reliable one-to-one comparison of the full benchmark setting and precluding threshold-based metrics such as AUC from the available aggregated outputs. Qualitatively, predictions associated with CN subjects appear concentrated within similar ranges, reflecting relatively consistent normal brain patterns, whereas predictions associated with AD span a broader range across the OASIS-1 cohort, indicating increased heterogeneity in pathological presentations. Although this analysis does not constitute a comprehensive fairness or demographic stratification study, it suggests that the main findings are not strictly dataset-specific. A systematic evaluation across larger and more diverse cohorts, such as AIBL or international datasets, remains an important direction for future work.

Acknowledgment of data source. Data were provided by the OASIS-1 Cross-Sectional dataset. Principal Investigators include D. Marcus, R. Buckner, J. Csernansky, and J. Morris. This work was supported by P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, and U24 RR021382.

7.2. Architectural design choices

Our slice-based 2D processing strategy represents a deliberate methodological trade-off. It enabled direct adaptation of pretrained 2D foundation encoders and allowed all modalities, fusion heads, and missing-modality settings to be studied within a unified framework. At the same time, this choice discards full volumetric context and cannot preserve z-axis spatial correlations, which are relevant for neurodegenerative patterns such as regional atrophy and ventricular enlargement. The 3D PET baselines reported earlier show that volumetric models can remain stronger when complete volumes are available, particularly in the binary PET-only setting. Accordingly, the present results should be interpreted as evidence for the utility of foundation-model-based multimodal analysis under a 2D slice formulation, rather than as evidence that 2D inputs are generally preferable to dedicated 3D architectures. Extending the framework to 3D foundation models or hybrid slice-volume strategies remains an important direction for future work.

7.3. Synthetic data generation and generative model limitations

Our ControlNet-based generative approach provides a practical strategy for handling missing imaging modalities in multimodal learning. By generating synthetic images that preserve coarse structural consistency, the method can support feature learning in settings with incomplete data. However, important limitations remain regarding biological fidelity and the reliability of the generated images.

First, the synthetic images have not undergone expert clinical validation, leaving uncertainty as to whether they accurately reflect realistic anatomical or pathological patterns. Moreover, the reconstruction metrics reported in Table 11 reveal consistently low SSIM and PSNR values across multiple modality pairs. These results indicate that, while global anatomical structure may be preserved, the generated images differ substantially from real target modalities at the intensity and local-structure level. Consequently, the synthetic data should not be interpreted as faithful biological reconstructions nor considered suitable for direct clinical interpretation.

These limitations introduce a concrete risk that generative augmentation may induce spurious correlations or modality-specific artifacts. Such effects could artificially improve internal validation metrics by reinforcing shortcut features, while negatively impacting generalization to unseen cohorts or external datasets. Our experimental findings are consistent with this risk, as performance gains from generative augmentation are observed only in specific configurations, whereas in other settings the inclusion of synthetic data degrades results, particularly when synthetic samples dominate over real images.

For these reasons, synthetic images are strictly confined to training and validation, and all reported test evaluations are performed exclusively on real data. The role of the generative model within our framework should therefore be understood as a limited training-time augmentation strategy for incomplete-modality settings. Its effect is configuration-dependent, can be harmful when synthetic inputs are used non-selectively, and should not be interpreted as a reliable tool for biologically accurate modality reconstruction.

7.4. Clinical deployment barriers

Despite the encouraging performance observed in our experimental evaluation, several barriers remain before the proposed framework

can be translated into routine clinical practice. One important limitation concerns interpretability. While the transformer-based fusion head relies on attention mechanisms that are, in principle, amenable to visualization, this study does not include a systematic analysis of attention maps or qualitative assessments by clinicians. Moreover, the internal representations learned by large transformer models remain difficult to directly relate to established neuroanatomical or pathological biomarkers, reinforcing their characterization as black-box systems. As such, attention-based explanations are not yet validated as reliable indicators of clinically meaningful decision factors, and their integration is left as a direction for future work. Another key challenge relates to real-world clinical workflows. Although the framework is explicitly designed to operate under incomplete multimodal settings by accommodating missing imaging modalities, this capability has been evaluated only in a retrospective, dataset-driven context. No user studies or prospective clinical evaluations were conducted to assess how generated or substituted modalities would be perceived, trusted, or acted upon by clinicians in practice. Consequently, while the proposed design offers a technically flexible solution to data incompleteness, its impact on clinical decision-making remains to be established. These considerations highlight that the present work should be viewed as a methodological contribution rather than a clinically validated system. Future studies will be required to assess interpretability tools in collaboration with domain experts and to evaluate the framework within realistic clinical workflows, including user feedback and prospective validation.

7.5. Methodological focus and baseline strategy

Our comparative analysis strategically emphasizes deep learning approaches to maintain methodological coherence and enable meaningful architectural comparisons. By focusing on foundation model adaptations and transformer-based fusion, we provide clear insights into the specific contributions of our approach within the modern AI paradigm. This focused comparison strategy, rather than exhaustive baseline coverage, allows for deeper investigation of our core hypotheses regarding multimodal fusion and generative augmentation. Future work can build upon our established foundation to explore hybrid approaches combining the strengths of different methodological paradigms. Accordingly, the current benchmark is not exhaustive with respect to all multimodal fusion and missing-modality design families, particularly branch-based cross-attention architectures and latent-space aggregation methods.

7.6. Fairness and representation

Our comprehensive fairness analysis reveals both the potential and challenges of applying AI to diverse populations. The observed performance variations across demographic groups, rather than representing a failure, highlight the critical importance of bias-aware model development and provide a baseline for future fairness-enhanced approaches. While this work provides these foundational insights, it does not implement explicit fairness-aware training or bias mitigation techniques, leaving the impact of such strategies to be explored in future studies. This analysis contributes valuable insights to the broader medical AI community regarding the need for inclusive dataset construction and algorithmic fairness in healthcare applications. Nevertheless, the reported subgroup analyses offer actionable insights that can inform future bias mitigation strategies. In particular, the observed imbalances suggest that oversampling underrepresented subgroups, reweighting loss functions based on subgroup prevalence, or incorporating domain adaptation mechanisms to reduce site-specific effects could further improve equity across populations. These directions are left for future work, where targeted mitigation techniques can be rigorously evaluated without conflating fairness assessment with intervention effects.

7.7. Translational opportunities

The limitations identified in this work collectively define a clear roadmap for clinical translation. The demonstrated effectiveness of our approach under controlled conditions provides strong motivation for prospective validation studies, integration with existing clinical workflows, and development of deployment-ready systems. Each limitation represents a specific research direction that can build upon our foundational contributions to advance the field toward practical clinical impact.

7.8. Temporal generalization

A further limitation of the present study concerns temporal generalization and longitudinal consistency. To reduce label ambiguity and avoid potential data leakage, our experimental design restricts the analysis to first-visit scans and excludes subjects whose diagnosis changes across visits. While this choice supports a cross-sectional, screening-oriented evaluation, it prevents a direct assessment of whether model predictions evolve consistently with disease progression over time. To partially explore this aspect, we performed an auxiliary evaluation on the subset of excluded patients exhibiting diagnostic progression, focusing on PET-based models trained using both clinical and synthetic data, since it was the best-performing combination across the MC approaches. On this challenging cohort, performance was markedly lower, with an F1 score of 0.4641 ± 0.2820 in the B setting and 0.2420 ± 0.1421 in the MC setting. The reduced accuracy and high variance indicate that progression cases are substantially more difficult and suggest that models trained under a cross-sectional assumption do not reliably capture longitudinal disease dynamics. A principled evaluation of temporal consistency would require dedicated longitudinal modeling strategies and multi-visit training protocols, which are beyond the scope of the present work.

8. Conclusion

This work shows that foundation models can serve as a useful basis for multimodal neuroimaging analysis under incomplete modality availability, while also making clear that neither multimodal fusion nor generative augmentation is uniformly beneficial across settings. Within this revised scope, the contribution of the manuscript is not to claim a universally superior multimodal solution, but to provide a unified and reproducible framework that clarifies where such strategies are competitive, where they degrade, and what limitations remain relative to strong unimodal volumetric baselines. Our comprehensive framework demonstrates that vision foundation models can be adapted to neuroimaging through Low-Rank Adaptation, achieving an F1-score of 57.80% in MC AD classification in the best multimodal setting with clinical integration and generative augmentation.

The ControlNet-based generative component provides a practical training-time augmentation strategy for incomplete multimodal settings. Its benefits were most evident in selected PET-centered configurations, whereas other settings showed limited gains or even degradation, indicating that synthetic modalities should be treated as auxiliary support for representation learning rather than as faithful replacements for missing clinical acquisitions.

This selective use of synthetic augmentation is relevant to real-world clinical constraints where cost, accessibility, or patient factors prevent comprehensive imaging protocols, but it should be treated as auxiliary support rather than a uniformly beneficial component. The transformer-based fusion architecture provides a flexible mechanism for modeling cross-modal interactions and was particularly advantageous in more complex multimodal configurations, although simpler aggregation remained competitive in several settings. Beyond technical contributions, our comprehensive fairness analysis provides a structured framework for performance and subgroup-level analysis for bias-aware medical AI development, revealing important performance variations across demographic groups while providing frameworks for equitable algorithm design. This dual focus on performance and fairness positions our work within the broader movement toward responsible healthcare AI. Our findings demonstrate that foundation models, originally developed for natural images, can be successfully adapted to specialized

medical domains while enabling a more cautious subgroup-level fairness assessment. This cross-domain transfer capability suggests promising applications across multiple medical imaging specialties, potentially accelerating AI development beyond neuroimaging. By bringing foundation model adaptation, multimodal fusion, and missing-modality augmentation into a single experimental framework, this work offers a methodological basis for future research on neuroimaging classification under incomplete modality availability. We hope these results help clarify which design choices are promising, which remain limited, and where stronger multimodal methods are still needed.

CRediT authorship contribution statement

Luca Zedda: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrea Loddo:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Formal analysis, Conceptualization. **Cecilia Di Ruberto:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001 - Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR), and by the GNCS Project 2025 “Metodi di approssimazione globale per operatori integrali e applicazioni alle equazioni funzionali” (CUP E53C24001950001).

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative1 (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.

Code and Data Availability

The code used for cohort filtering, preprocessing, training, and evaluation is publicly available at https://github.com/unica-visual-intelligence-lab/Multimodal-Neuroimaging. The study uses ADNI data obtained from the LONI Image Data Archive and the OASIS-1 dataset, which remain subject to their respective access and usage conditions. In accordance with these source-dataset terms, the manuscript does not redistribute original imaging data or derived synthetic images.

Data availability

The authors do not have permission to share data.

References

[1] M.A. Azam, K.B. Khan, S. Salahuddin, E. Rehman, S.A. Khan, M.A. Khan, S. Kadry, A.H. Gandomi, A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, Comput. Biol. Med. 144 (2022) 105253, https://doi.org/10.1016/j.compbiomed.2022.105253, https://www.sciencedirect.com/science/article/pii/S0010482522000452.

[2] I.U. Haq, M. Mhamed, M. Al-Harbi, H. Osman, Z.Y. Hamd, Z. Liu, Advancements in medical radiology through multimodal machine learning: a comprehensive overview, Bioengineering 12 (5) (2025) 477, https://doi.org/10.3390/bioengineering12050477, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12108733/.

[3] H.A. Helaly, M. Badawy, A.Y. Haikal, Deep learning approach for early detection of alzheimer’s disease, Cogn. Comput. 14 (5) (2021) 1711–1727, Publisher: Springer. https://doi.org/10.1007/s12559-021-09946-2, https://link.springer.com/article/10.1007/s12559-021-09946-2.

[4] V.S. Diogo, H.A. Ferreira, D. Prata, Early diagnosis of alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach, Alzheimer’s Res. Ther. 14 (1) (2022) 1–21, Publisher: BioMed Central. https://doi.org/10.1186/s13195-022-01047-y, https://link.springer.com/article/10.1186/s13195-022-01047-y.

[5] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of alzheimer’s disease: overview and reproducible evaluation, Med. Image Anal. 63 (2020) 101694, https://doi.org/10.1016/j.media.2020.101694, https://www.sciencedirect.com/science/article/pii/S1361841520300591.

[6] B. Hou, Z. Wang, Z. Zhou, B. Tong, Z. Wang, J. Bao, D. Duong-Tran, Q. Long, L. Shen, Fair CCA for Fair Representation Learning: an ADNI Study, Association for Computing Machinery, New York, NY, USA, 2025, https://doi.org/10.1145/3765612.3767215

[7] G. Mulugeta, M.A. Eckert, K.I. Vaden, T.D. Johnson, A.B. Lawson, Methods for the analysis of missing data in FMRI studies, Journal of biometrics & biostatistics 8 (1) (2017) 335, https://doi.org/10.4172/2155-6180.1000335, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6510494/.

[8] M. Liu, S. Li, H. Yuan, M.E.H. Ong, Y. Ning, F. Xie, S.E. Saffari, Y. Shang, V. Volovici, B. Chakraborty, N. Liu, Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques, Artif. Intell. Med. 142 (2023) 102587, https://doi.org/10.1016/j.artmed.2023.102587, https://www.sciencedirect.com/science/article/pii/S093336572300101X.

[9] O. Siméoni, H.V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski, DINOv3, cs, Aug 2025 arXiv:2508.10104.

[10] L. Zedda, A. Loddo, C. Di Ruberto, Radio DINO: a foundation model for advanced radiomics and AI-driven medical imaging analysis, Comput. Biol. Med. 195 (2025) 110583, https://doi.org/10.1016/j.compbiomed.2025.110583, https://www.sciencedirect.com/science/article/pii/S0010482525009345.

[11] T. Vanaja, K. Shanmugavadivel, M. Subramanian, C.S. Kanimozhiselvi, Advancing alzheimer’s detection: integrative approaches in MRI analysis with traditional and deep learning models, Neural Comput. Appl. 37 (14) (2025) 8527–8546, Publisher: Springer. https://doi.org/10.1007/s00521-025-10993-1, https://link.springer.com/article/10.1007/s00521-025-10993-1.

[12] I. Jahani, A. Jahani, M. Delrobaei, A. Khadem, B.J. MacIntosh, Classifying cognitive impairment based on FDG-PET and combined t1-MRI and rs-fMRI: an ADNI study, J. Alzheimer’s Dis. 103 (2) (2025) 452–464, https://doi.org/10.1177/13872877241302493, Publisher: SAGE Publications.

[13] R. Karakis, K. Gurkahraman, E. Unsal, B. Cigdem, Detection of alzheimer’s disease with an ensemble deep learning model using diffusion tensor imaging, in: Futuristic Computational Systems and Advanced Engineering for the Society, 2025, pp. 25–37, https://doi.org/10.1007/978-3-031-92552-8_3, https://link.springer.com/chapter/10.1007/978-3-031-92552-8_3.

[14] Z. Ur-Rehman, M.K. Awang, G. Ali, M. Faheem, Deep learning techniques for alzheimer’s disease detection in 3d imaging: a systematic review, Health Sci. Rep. 7 (9) (2024) e70025, _eprint: https://doi.org/10.1002/hsr2.70025.

[15] I. Malik, A. Iqbal, Y.H. Gu, M.A. Al-Antari, Deep learning for alzheimer’s disease prediction: a comprehensive review, Diagnostics 14 (12) (2024) 1281, https://doi.org/10.3390/diagnostics14121281, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11202897/.

[16] A. Nazir, A. Assad, A. Hussain, M. Singh, Alzheimer's disease diagnosis using deep learning techniques: datasets, challenges, research gaps and future directions, Int. J. Syst. Assur. Eng. Manag. (2024) 1–35Publisher: Springer. https://doi.org/10.1007/s13198-024-02441-5, https://link.springer.com/article/10.1007/s13198-024-02441-5.

[17] S.K. Hammonds, T. Eftestøl, K. Oppedal, A. Fernandez-Quilez, Unraveling gender fairness analysis in deep learning prediction of alzheimer's disease, in: 2024 4th International Conference on Applied Artificial Intelligence (ICAPAI), 2024, pp. 1–7, https://doi.org/10.1109/ICAPAI61893.2024.10541140, https://ieeexplore.ieee.org/abstract/document/10541140.

[18] K. Zhang, Promoting equity: assessing algorithmic fairness in machine learning approaches for predicting alzheimer's disease, in: 2024 IEEE International Conference on future Machine Learning and Data Science (FMLDS), 2024, pp. 315–318, https://doi.org/10.1109/FMLDS63805.2024.00063, https://ieeexplore.ieee.org/abstract/document/10874063.

[19] B. Tong, T. Edwards, S. Yang, B. Hou, D.A. Tarzanagh, R.J. Urbanowicz, J.H. Moore, M.D. Ritchie, C. Davatzikos, L. Shen, Ensuring fairness in detecting mild cognitive impairment with MRI, AMIA Annual Symposium Proceedings 2024 (2025) 1119–1128, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12099326/.

[20] B. Li, X. Jiang, K. Zhang, A.O. Harmanci, B. Malin, H. Gao, X. Shi, T.A.D.N. Initiative, Enhancing fairness in disease prediction by optimizing multiple domain adversarial networks, PLOS Digit. Health (2025), https://doi.org/10.1371/journal.pdig.0000830, https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000830.

[21] R. Mehta, C. Shui, T. Arbel, Evaluating the fairness of deep learning uncertainty estimates in medical image analysis, in: I. Oguz, J. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heinmann, D. Kontos, B. Landman, B. Dawant (Eds.), Medical Imaging with Deep Learning, 227, PMLR, 2024, pp. 1453–1492, https://proceedings.mlr.press/v227/mehta24a.html. Proceedings of Machine Learning Research.

[22] S.W. Park, N.Y. Yeo, Y. Kim, G. Byeon, J.-W. Jang, Deep learning application for the classification of alzheimer's disease using 18f-flortaucipir (AV-1451) tau positron emission tomography, Sci. Rep. 13 (1) (2023) 8096, https://doi.org/10.1038/s41598-023-35389-w, https://www.nature.com/articles/s41598-023-35389-w.

[23] for the Alzheimer's Neuroimaging Initiative, T. Jo, K. Nho, S.L. Risacher, A.J. Saykin, Deep learning detection of informative features in tau PET for alzheimer's disease classification, BMC Bioinform. 21 (S21) (2020) 496, https://doi.org/10.1186/s12859-020-03848-0, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03848-0.

[24] M.M.S. Fareed, S. Zikria, G. Ahmed, Mui-Zzud-Din, S. Mahmood, M. Aslam, S.F. Jillani, A. Moustafa, M. Asad, ADD-net: an effective deep learning model for early detection of Alzheimer disease in MRI scans, IEEE Access 10 (2022) 96930–96951, https://doi.org/10.1109/ACCESS.2022.3204395, https://ieeexplore.ieee.org/abstract/document/9877809.

[25] P. Saikia, S.K. Kalita, Alzheimer disease detection using MRI: deep learning review, SN Comput. Sci. 5 (5) (2024) 1–16, Publisher: Springer. https://doi.org/10.1007/s42979-024-02868-4, https://link.springer.com/article/10.1007/s42979-024-02868-4.

[26] V. Zarovniaeva, S. Anwar, S. Kazmi, K. Cortez Perez, S. Sandhu, L. Mohammed, The role of PET detection of biomarkers in early diagnosis, progression, and prognosis of alzheimer's disease: a systematic review, Cureus 17 (1) (2025) e77781, https://doi.org/10.7759/cureus.77781, https://pmc.ncbi.nlm.nih.gov/articles/PMC11841692/.

[27] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, G. Catheline, 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies, cs, Jan 2018, arXiv:1801.05968.

[28] A. Tiwari, A. Singhal, S.J. Shigwan, R.K. Singh, Early diagnosis of Alzheimer through Swin-Transformer-Based deep learning framework using sparse diffusion measures, in: Proceedings of the 15th Asian Conference on Machine Learning, 2024, pp. 1369–1384, https://proceedings.mlr.press/v222/tiwari24a.html.

[29] S. Qiu, M.I. Miller, P.S. Joshi, J.C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P.H. Hwang, J.A. Cramer, B.C. Dwyer, H. Hao, M.C. Kaku, S. Kedar, P.H. Lee, A.Z. Mian, D.L. Murman, S. O'Shea, A.B. Paul, M.-H. Saint-Hilaire, E. Alton Sartor, A.R. Saxena, L.C. Shih, J.E. Small, M.J. Smith, A. Swaminathan, C.E. Takahashi, O. Taraschenko, H. You, J. Yuan, Y. Zhou, S. Zhu, M.L. Alosco, J. Mez, T.D. Stein, K.L. Poston, R. Au, V.B. Kolachalama, Multimodal deep learning for alzheimer's disease dementia assessment, Nat. Commun. 13 (1) (2022) 3404, https://doi.org/10.1038/s41467-022-31037-5, https://www.nature.com/articles/s41467-022-31037-5.

[30] S. Sharma, P.K. Mandal, A comprehensive report on machine learning-based early detection of alzheimer's disease using multi-modal neuroimaging data, ACM Comput. Surv. 55 (2) (2023) 1–44, https://doi.org/10.1145/3492865, https://dl.acm.org/doi/10.1145/3492865.

[31] J. Venugopalan, L. Tong, H.R. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of alzheimer's disease stage, Sci. Rep. 11 (1) (2021) 3254, Publisher: Nature Publishing Group. https://doi.org/10.1038/s41598-020-74399-w, https://www.nature.com/articles/s41598-020-74399-w.

[32] Y. Zhao, Q. Guo, Y. Zhang, J. Zheng, Y. Yang, X. Du, H. Feng, S. Zhang, Application of deep learning for prediction of alzheimer's disease in PET/MR imaging, Bioengineering 10 (10) (2023) 1120, Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/bioengineering10101120, https://www.mdpi.com/2306-5354/10/10/1120.

[33] M. Golovanevsky, C. Eickhoff, R. Singh, Multimodal attention-based deep learning for alzheimer's disease diagnosis, J. Am. Med. Inform. Assoc. 29 (12) (2022) 2014–2022, https://doi.org/10.1093/jamia/ocac168

[34] M. Aghili, S. Tabarestani, M. Adjouadi, Addressing the missing data challenge in multi-modal datasets for the diagnosis of alzheimer's disease, J. Neurosci. Methods 375 (2022) 109582, https://doi.org/10.1016/j.jneumeth.2022.109582, https://www.sciencedirect.com/science/article/pii/S0165027022001091.

[35] K.-H. Thung, P.-T. Yap, D. Shen, Multi-stage diagnosis of alzheimer's disease with incomplete multimodal data via multi-task deep learning, in: -, 2017, pp. 1–17, https://doi.org/10.1007/978-3-319-67558-9_19, https://link.springer.com/chapter/10.1007/978-3-319-67558-9_19.

[36] R. Azad, M. Dehghanmanshadi, N. Khosravi, J. Cohen-Adad, D. Merhof, Addressing missing modality challenges in MRI images: a comprehensive review, Comput. Vis. Media 11 (2) (2025) 241–268, https://doi.org/10.26599/CVM.2025.9450399, https://ieeexplore.ieee.org/abstract/document/10984423.

[37] M.I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, D. Hjelm, Mutual information neural estimation, in: Proceedings of the 35th International Conference on Machine Learning, 2018, https://proceedings.mlr.press/v80/belghazi18a.html.

[38] T. Sylvain, F. Dutil, T. Berthier, L. Di Jorio, M. Luck, D. Hjelm, Y. Bengio, CMIM: Cross-Modal information maximization for medical imaging, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 1190–1194, ISSN: 2379-190X. https://doi.org/10.1109/ICASSP39728.2021.9414132, https://ieeexplore.ieee.org/document/9414132/.

[39] M. Abdelaziz, T. Wang, A. Elazab, Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks, J. Biomed. Inform. 121 (2021) 103863, https://doi.org/10.1016/j.jbi.2021.103863, https://www.sciencedirect.com/science/article/pii/S1532046421001921.

[40] G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, Predicting alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (1) (2019) 1952, Publisher: Nature Publishing Group. https://doi.org/10.1038/s41598-018-37769-z, https://www.nature.com/articles/s41598-018-37769-z.

[41] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, HeMIS: Hetero-Modal Image Segmentation, 2016, https://doi.org/10.1007/978-3-319-46723-8_54

[42] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, J. Zhao, M3care: learning with missing modalities in multimodal healthcare data, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2418–2428, https://doi.org/10.1145/3534678.3539388, KDD '22.

[43] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, cs, Oct 2021, arXiv:2106.09685.

[44] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to Text-to-Image diffusion models, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, 2023, pp. 3813–3824, https://doi.org/10.1109/ICCV51070.2023.00355, https://ieeexplore.ieee.org/document/10377881/.

[45] RadImageNet, An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning, 2022, https://doi.org/10.1148/ryai.210315

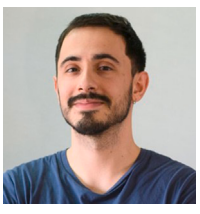
[46] A. Loddo, S. Buttau, C. Di Ruberto, Deep learning based pipelines for alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method, Comput. Biol. Med. 141 (2022) 105032, https://doi.org/10.1016/j.compbiomed.2021.105032, https://www.sciencedirect.com/science/article/pii/S001048252100826X.

[47] S. Chen, K. Ma, Y. Zheng, Med3D: Transfer Learning for 3D Medical Image Analysis, cs, Jul 2019, arXiv:1904.00625.

[48] Y. Tang, D. Yang, W. Li, H.R. Roth, B.A. Landman, D. Xu, V. Nath, A. Hatamizadeh, Self-Supervised Pre-Training of swin transformers for 3d medical image analysis, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 20698–20708.

[49] D.S. Marcus, T.H. Wang, J. Parker, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, J. Cogn. Neurosci. 19 (9) (2007) 1498–1507. arXiv: https://doi.org/10.1162/jocn.2007.19.9.1498, https://direct.mit.edu/jocn/article-pdf/19/9/1498/1936514/jocn.2007.19.9.1498.pdf.

Author biography



Luca Zedda is a Ph.D. student in Computer Science at the University of Cagliari, specializing in artificial intelligence and its applications in biomedical and clinical research. His work spans foundational topics in deep learning, computer vision, and machine learning, with a strong focus on medical imaging, radiomics, and high-dimensional biomedical data analysis. He is particularly interested in advancing self-supervised learning techniques, transformer architectures, and foundation models to address challenges in healthcare and life sciences. Luca's research aims to develop state-of-the-art AI systems that enhance disease detection, diagnosis, and data-driven healthcare solutions. He has contributed to various scientific publications and presented at international conferences, reflecting his commitment to pushing the boundaries of AI-driven innovations in scientific and clinical contexts.

Andrea Loddo received the B.Sc., M.Sc., and Ph.D. degrees from the University of Cagliari, in 2012, 2014, and 2019, respectively. His Ph.D. thesis addressed blood cell image analysis and classification issues to create new tools for automatic diagnosis to support medical analysis. He is currently an Assistant Professor at the Department of Mathematics and Computer Science, University of Cagliari. He has authored over 50 scientific manuscripts in peer-reviewed journals and international conference proceedings. His research interests include image analysis and processing, computer vision, pattern recognition, and machine and deep learning, with a focus on medical tasks.

Cecilia Di Ruberto is an Associate Professor of Computer Science at the Department of Mathematics and Computer Science, University of Cagliari, Italy. She received the M.Sc. in Computer Science from the University of Salerno, Italy, in 1990 and the Ph.D. degree in Computer Science from the University of Naples, Italy, in 1995. Currently, her research interests include computer vision, image retrieval, medical image analysis, pattern recognition, and machine learning. She has been working on microscopic image analysis, particularly in blood smear image analysis for cell counting, malaria parasite detection and classification, and leukemia detection. She is the author of over 100 scientific papers in peer-reviewed journals and international conference proceedings.