Multimodal Deep Learning Methods for Alzheimer's Disease Diagnosis

by

Mattia Cestari

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,
to be defended publicly on Monday July 20th, 2026 at 04:00 PM.

Student number:	6296068	
Project duration:	November 15th, 2025 – July 20th, 2026	
Thesis committee:	H. Kekkonen	TU Delft, supervisor
	Ö. Şahin	TU Delft
	A. Bhat	University of Chicago, supervisor

TU Delft THE UNIVERSITY OF CHICAGO

Acknowledgements

I would like to thank my TU Delft supervisor, Hanne, for her guidance and advice during these months. My deepest thanks also go to Avinay, for supporting me throughout this journey, and to Prof. Antonio Ereditato, for sponsoring countless hours of computation on the RCC. I am very grateful for the trust you have placed in me. This work would not have been possible without the ONAOSI Foundation and the University of Chicago, and the collaboration between them has been an incredible opportunity for learning and personal growth. Finally, I would like to thank my family, who are always caring, supportive and sweet, and my dear friends scattered throughout Italy, Europe and the world.

Mattia Cestari
Delft, July 2026

Abstract

Alzheimer's disease is the leading cause of dementia worldwide, and its accurate diagnosis is essential for the effectiveness of the therapies currently available. Deep learning methods applied to neuroimaging data have shown promise for this task, but most multimodal approaches assume that all modalities are available for every subject, an assumption that rarely holds in clinical practice or even in research cohorts such as ADNI, where FDG-PET is far less common than MRI.

This thesis develops a multimodal framework for the three-class classification of subjects into cognitively normal, mild cognitive impairment and Alzheimer's disease, which is natively robust to missing modalities. We formulate a Product-of-Experts model in which MRI and FDG-PET each contribute a Gaussian expert to a shared latent representation of the brain state, allowing predictions to be made from either modality alone or from both, always through the same classifier and without imputing the missing modality. Since different modality-availability patterns induce different latent-input regimes for this shared classifier, we further investigate two regularization strategies: a meta-learned regularizer, inspired by domain-generalization methods, and a Kullback-Leibler term that encourages the unimodal and multimodal latent distributions to remain compatible.

Experiments on ADNI data show that jointly training the model with a standard supervised objective does not by itself improve over dedicated single-modality baselines, and that the latent representations induced by different modality subsets are clearly separated. Both regularizers improve on this basic model, with the meta-learned regularizer providing the most consistent gains across the three inference pathways. In the multimodal setting, the proposed model reaches a balanced accuracy comparable to that of a recent state-of-the-art transformer-based method, while using roughly two orders of magnitude fewer trainable parameters and remaining usable when one modality is absent at inference time.

Contents

Acknowledgements i

Abstract ii

1 Introduction 1

2 Background and related work 3
- 2.1 Alzheimer's disease . . . 3
- 2.2 Alzheimer's Disease Neuroimaging Initiative . . . 4
- 2.3 Neuroimaging modalities . . . 5
  - 2.3.1 Structural MRI . . . 5
  - 2.3.2 FDG-PET . . . 6
- 2.4 Machine learning for Alzheimer's disease classification . . . 7
  - 2.4.1 Multimodal learning . . . 8
  - 2.4.2 Evaluation issues in ADNI . . . 8
  - 2.4.3 Positioning of the present thesis . . . 8

3 Probabilistic model 10
- 3.1 Motivation . . . 10
- 3.2 Shared latent representation . . . 12
- 3.3 Product of Experts . . . 13
- 3.4 Parametrization . . . 15
  - 3.4.1 Gaussian parametrization . . . 15
  - 3.4.2 MVAE-Q reparametrization . . . 16
- 3.5 Final model . . . 18

4 Meta-learning training 20
- 4.1 Motivation . . . 20
- 4.2 Meta-learning regularization for cross-regime generalization . . . 22
  - 4.2.1 Meta-learning for domain generalization . . . 22
  - 4.2.2 Meta-regularization . . . 23
  - 4.2.3 Adaptation to the Product-of-Experts model . . . 24
- 4.3 KL regularization . . . 25
- 4.4 Training algorithms . . . 27

5 Experimental Setup 30
- 5.1 Dataset and Preprocessing . . . 30
  - 5.1.1 ADNI original data . . . 30
  - 5.1.2 MRI preprocessing . . . 31
  - 5.1.3 FDG-PET preprocessing . . . 31
  - 5.1.4 Label matching . . . 33
  - 5.1.5 Multimodal samples creation . . . 33
  - 5.1.6 Final dataset . . . 35
  - 5.1.7 Training, validation and test splits . . . 35
  - 5.1.8 Augmentation . . . 37
- 5.2 Model architecture . . . 39
  - 5.2.1 Modality-specific encoders . . . 39
  - 5.2.2 Shared classifier . . . 41
- 5.3 Stochastic and Deterministic objective . . . 42
- 5.4 Evaluation Metrics . . . 43
- 5.5 Experiment design . . . 44

5.5.1 Model selection procedure . . . 45
5.5.2 Stochastic vs Deterministic approximation and Baseline experiments . . . 45
5.5.3 Mini-batch construction for the joint objective . . . 45
5.5.4 MetaReg and KL regularization experiments . . . 46

6 Results **48**
6.1 Stochastic vs Deterministic approximations and Baseline experiments . . . 48
6.2 Basic Product-of-Experts . . . 51
6.3 Product-of-Experts with KL regularization . . . 53
6.4 Product-of-Experts with MetaReg . . . 55
6.5 Product-of-Experts with KL and MetaReg regularization . . . 56

7 Conclusion **58**
7.1 Main findings . . . 58
7.2 Comparison with the state of the art . . . 59
7.3 Limitations . . . 60
7.4 Future work . . . 60

Bibliography **61**

1

Introduction

Alzheimer's disease is the leading cause of dementia worldwide, expected to affect over 130 million people globally by 2050 with an estimated economic impact of $ 2 trillion. It is a neurodegenerative disorder that gradually impairs the patient's cognitive abilities, initially by affecting memory and orientation, then corrupting social functioning and interactions, and finally leading to the inability to perform basic voluntary actions, which ultimately provokes death.

To the current date, the pathogenesis of Alzheimer's disease is still heavily debated and not fully understood, partly due to the complexity of the biological mechanisms that are shared by a variety of neurological disorders. Even though this lack of understanding has considerably hindered the development of proper disease-modifying drugs, to date, a few therapies have demonstrated modest efficacy in slowing disease progression. However, the success of such therapies heavily depends on the ability to diagnose the stage of the disease timely with high accuracy, in order to begin treatment as early as possible.

Over the years, several evaluation scales and frameworks have been developed with the aim of diagnosing the disease stage, which can be classified as cognitively normal (CN), mild cognitive impairment (MCI), a transitional stage, or Alzheimer's disease (AD). These methods mostly rely on the assessment of the clinical symptoms which are however common to other types of dementia and disorders; because of this, these diagnosis methods lack high specificity to AD, which is required for AD-specific treatments as they may not otherwise be effective against other types of dementia. Moreover, a greater challenge consists in the early diagnosis of AD, which would allow to begin therapies even before symptoms have arisen, during the preclinical stage.

These core challenges motivated the Alzheimer's Disease Neuroimaging Initiative (ADNI), which set out in 2005 with the purpose of investigating possible biomarkers of the disease, i.e. quantifiable biological changes caused by the pathogenic process, which could then serve as more accurate and possibly early diagnostic tools. As of today, this longitudinal study is still running and, over the course of more than 20 years, it has gathered data of over 3300 patients from U.S. and Canada. As the name suggests, the study focuses primarily on the collection of neuroimaging data, in particular Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans of the subjects' brains. The availability of such neuroimaging data and advances in machine learning have sparked interest in the application of these models to the tasks of disease stage classification and early diagnosis.

In particular, the classification of subjects into the three diagnostic categories, cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD), has emerged as a central benchmark for evaluating multimodal deep learning architectures. While it is important to acknowledge that clinical diagnosis of AD is inherently multimodal and longitudinal, integrating cognitive assessments, fluid biomarkers, and imaging over time, the CN, MCI, and AD classification task using single timepoint neuroimaging serves as a well defined and publicly available testbed.

Recent work has pushed the state of the art on this benchmark. Notably, DiaMond [1] employs a Vision Transformer architecture to fuse MRI and FDG-PET, achieving a balanced accuracy of 65.2% and a macro F1 score of 64.9%. Although there appears to be a large margin for improvement, these metrics represent the state of the art as the three-way CN vs MCI vs AD classification task is notoriously difficult, partly due to the lack of a sufficiently large dataset, partly because the MCI class is heterogeneous, making it hard to recognize. Despite its strong performance, DiaMond exhibits two limitations common in the literature. First, it assumes that both modalities are available for every subject at both training and inference time, an assumption that rarely holds in clinical practice or even within ADNI itself, where many subjects lack one or more modalities. Second, its transformer-based design is computationally heavy: instantiating the publicly available implementation with the configuration adopted in [1] yields approximately 30 million trainable parameters, potentially hindering deployment in resource-constrained settings.

This thesis addresses these gaps by developing a multimodal deep learning framework that is natively robust to missing modalities and efficient in its parameterization. Specifically, we pursue three primary objectives:

1. We formulate a Product-of-Experts model that infers a shared latent representation of brain state from any available subset of modalities, namely MRI, PET, or both. This probabilistic framework naturally handles missing modalities at inference time without requiring modality imputation.
2. We develop a meta-learning based training strategy inspired by domain-generalization methods such as MLDG [2] and by meta-learned regularization [3]. The goal is to improve robustness across the latent-input regimes induced by different modality-availability patterns. In addition, we introduce a KL-based alignment term that encourages compatibility between unimodal and multimodal distributions.
3. We compare the performance of the proposed framework to unimodal baselines, standard multimodal training, and recent methods for multimodal Alzheimer's disease classification [1]. Further, we perform ablation studies to evaluate the contributions the different regularization techniques.

The remainder of this thesis is structured as follows. Chapter 2 provides background on Alzheimer's disease, neuroimaging biomarkers, the ADNI study, and related work on deep learning methods for AD classification. Chapter 3 develops the probabilistic foundations of our approach, while Chapter 4 presents the meta-learning training framework and the KL-based latent alignment strategy. Chapter 5 discusses data preprocessing, the model architecture and the experiment design. Chapter 6 reports the results, and Chapter 7 discusses their implications, limitations, and directions for future work.

2

Background and related work

This chapter provides the background needed for the rest of this thesis. Section 2.1 introduces Alzheimer's disease, its clinical stages and its pathological hallmarks. Section 2.2 describes the Alzheimer's Disease Neuroimaging Initiative and its role in the search for disease biomarkers. Section 2.3 presents the two neuroimaging modalities used in this thesis, structural MRI and FDG-PET, along with their respective strengths and limitations. Section 2.4 reviews machine learning approaches to Alzheimer's disease classification, discusses the specific challenges of multimodal learning and of evaluation on ADNI data, and concludes by positioning the present work within this body of literature.

2.1. Alzheimer's disease

The first known case of Alzheimer's disease was documented in 1906, when Alois Alzheimer presented the clinical case of Auguste Deter, a patient who had developed severe memory loss, paranoia, disorientation and hallucinations before dying at the age of 55 [4]. After her death, Alzheimer examined her brain and identified abnormal protein deposits, later recognized as extracellular plaques and intracellular neurofibrillary tangles [5]. Subsequent advances, particularly immunohistochemistry, showed that these deposits are mainly composed of amyloid and tau proteins, respectively [5, 6]. These amyloid plaques and tau tangles are now considered the defining pathological hallmarks of AD, although definitive diagnosis historically required post-mortem confirmation [4].

For several decades, AD was regarded as a rare form of pre-senile dementia affecting individuals between 50 and 65 years of age, while cognitive decline in older patients was often interpreted as senile dementia and considered part of normal ageing. This view changed around 1970, when studies showed that patients with senile dementia exhibited the same pathological hallmarks as those with pre-senile dementia [4]. This led to the unification of these conditions under the term Alzheimer's disease, with early-onset and late-onset forms distinguished mainly by age of onset [7, 8]. As a consequence, AD became recognized as a major public health issue and a leading cause of death, motivating extensive research into its diagnosis, pathophysiology and treatment [7, 9].

Clinically, AD is a progressive neurodegenerative disorder that gradually impairs cognitive abilities. The National Institute on Aging and Alzheimer's Association framework describes three main stages [10]. The first is the preclinical stage, in which no clinical symptoms are present. This phase is also referred to as cognitively normal (CN) [11]. The second stage is mild cognitive impairment (MCI), or early AD, characterized by decline in memory, attention, language and visuospatial skills. Since these symptoms overlap with those of other neurological disorders, diagnosis must exclude alternative causes such as Parkinsonism, cerebrovascular risk factors or hallucinations [10]. The final stage is AD dementia, which may be moderate or severe. Moderate

AD involves behavioral changes, worsening memory loss and difficulty recognizing familiar people, while severe AD leads to loss of communication and disruption of functions such as swallowing, walking and urination, ultimately leading to death [10].

AD is currently the leading cause of dementia worldwide, accounting for approximately 60–70% of all dementia cases. In 2021, about 57 million people were affected, with around 10 million new cases each year [12]. Forecasts suggest that this number may reach 132 million by 2050, partly due to increased life expectancy and population ageing, with global healthcare and medical costs estimated at around $2 trillion [13].

Despite the long-standing identification of amyloid plaques and tau tangles, the biological mechanisms causing progressive cognitive decline remain unclear and debated [14]. The amyloid-cascade hypothesis, proposed in 1992, identified amyloid accumulation as the main causal factor in AD pathogenesis [15]. Although this hypothesis has been refined over time, increasing evidence points to a more complex disease process involving multiple interacting mechanisms and trajectories shared with other neuropathologies [16]. Moreover, clinical trials targeting amyloid deposition have generally produced limited results, with only a few reporting slowed disease progression, thereby challenging a purely amyloid-centered explanation of AD [14]. These unresolved questions have motivated increasing interest in biomarkers: measurable indicators of pathological processes that may support earlier and more accurate characterization of the disease. This motivation led to the development of the Alzheimer's Disease Neuroimaging Initiative, a large-scale biomarker study which is discussed in the following section.

2.2. Alzheimer's Disease Neuroimaging Initiative

The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2005 as a large, multi-site longitudinal study designed to support the discovery and validation of biomarkers for Alzheimer's disease. The study was motivated by the need for diagnostic tools that could identify AD-related pathology more accurately than clinical symptoms alone, detect the disease at earlier stages, and monitor progression over time. This objective became particularly relevant in the context of disease-modifying therapies, since treatments targeting AD-specific mechanisms require reliable identification of patients whose cognitive decline is actually related to Alzheimer's pathology [17].

A biomarker is a measurable biological indicator providing information that is closely related to the disease process. Ideally, it should reflect a core pathological mechanism of AD, distinguish AD from other causes of dementia, correlate with disease stage and cognitive decline, and be measurable in a reliable, minimally invasive and affordable way. ADNI was created precisely to study such biomarkers in a standardized setting, combining clinical assessments, neuroimaging, genetic information, fluid biomarkers and longitudinal follow-up [17].

Neuroimaging has played a central role in ADNI. Historically, structural imaging was mainly used to exclude alternative causes of cognitive impairment, such as tumors, vascular lesions or other abnormalities unrelated to AD. Over time, however, neuroimaging became increasingly important not only as an exclusion criterion, but also as a source of quantitative biomarkers. Structural Magnetic Resonance Imaging (MRI) can reveal patterns of brain atrophy associated with neurodegeneration, while Positron Emission Tomography (PET) can provide functional or molecular information depending on the radiotracer used. In ADNI, MRI and PET therefore represent two complementary views of the disease process.

The first phase of ADNI enrolled cognitively normal subjects, patients with mild cognitive impairment, and patients with Alzheimer's disease. Participants were followed longitudinally, with repeated clinical evaluations and imaging acquisitions. Subsequent phases, including ADNI-GO, ADNI2, ADNI3 and ADNI4, extended the original study by enlarging the cohort, refining acquisition protocols and introducing new biomarkers. Across these phases, the emphasis of PET imaging also evolved. FDG-PET was prominent in the early phases because of its ability to measure cerebral glucose metabolism, while later phases increasingly included amyloid and tau PET tracers, which more directly target pathological hallmarks of AD. Nevertheless, FDG-PET

remains highly relevant for machine learning studies because it is one of the more widely available PET modalities in ADNI and provides strong diagnostic information [17].

For the purposes of this thesis, ADNI is particularly important because it provides a public and extensively studied dataset for the classification of subjects into cognitively normal, mild cognitive impairment and Alzheimer's disease groups. At the same time, ADNI also reflects a central practical difficulty of multimodal medical data: not all subjects have all modalities available. MRI is more common, whereas PET is more limited because of cost, invasiveness and logistical constraints. This naturally motivates methods that can learn from both complete MRI and PET samples and incomplete MRI-only samples.

2.3. Neuroimaging modalities

2.3.1. Structural MRI

Magnetic Resonance Imaging is a non-invasive structural imaging technique that produces detailed images of brain anatomy. An example of a slice from an acquisition volume is shown in Figure 2.1. In the context of Alzheimer's disease, structural MRI is mainly used to measure patterns of neurodegeneration, such as cortical thinning, ventricular enlargement and regional atrophy. Particular attention is often given to medial temporal structures, including the hippocampus, because atrophy in these regions is strongly associated with memory impairment and disease progression [18].

MRI is based on the magnetic properties of hydrogen nuclei in water molecules. When the subject is placed inside a strong magnetic field, these nuclei become partially aligned with the field. Radiofrequency pulses then perturb this alignment, and the subsequent relaxation of the nuclei generates signals that can be measured by the scanner. By varying acquisition parameters, different tissue contrasts can be obtained. T1-weighted MRI, which is widely used in ADNI and in structural neuroimaging studies, provides good contrast between gray matter, white matter and cerebrospinal fluid, making it suitable for anatomical analysis [19].

In ADNI, MRI scans have been acquired using different scanners and acquisition protocols, including both 1.5T and 3T field strengths. Higher field strengths generally provide increased signal-to-noise ratio and potentially better spatial resolution, but they may also introduce differences in contrast and image characteristics [20]. In this thesis, both 1.5T and 3T scans are treated as belonging to the same MRI modality. This choice increases the amount of usable data, but it also introduces additional heterogeneity, which should be considered as a possible limitation.

From a machine learning perspective, MRI is attractive because it is widely available, as it is non-invasive and relatively inexpensive compared to PET. It also provides a stable structural representation of the brain, allowing models to learn disease-related anatomical patterns. However, MRI captures only one aspect of the disease process. Structural atrophy is informative, but it may appear relatively late compared with molecular or metabolic changes. This motivates the use of complementary modalities such as FDG-PET.

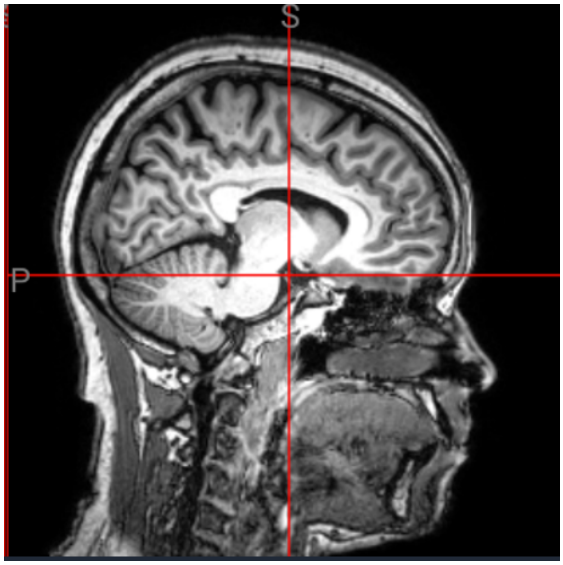


Figure 2.1: Sagittal slice of an MRI volume from the ADNI dataset

2.3.2. FDG-PET

Positron Emission Tomography is a functional imaging technique that measures the spatial distribution of a radioactive tracer injected into the subject. In FDG-PET, the tracer is fluorodeoxyglucose with fluorine-18, commonly denoted by [^{18}F]FDG. Since FDG is an analogue of glucose, it is taken up by metabolically active cells. After uptake, it becomes trapped in the tissue long enough for its radioactive decay to be detected by the PET scanner. The resulting image reflects regional glucose metabolism [21]. An example slice from a FDG-PET acquisition is shown in Figure 2.2.

In Alzheimer's disease, FDG-PET is informative because neurodegeneration is associated with characteristic patterns of hypometabolism. Reduced glucose metabolism in temporoparietal and related cortical regions has been linked to synaptic dysfunction and cognitive decline. Therefore, whereas MRI reflects structural damage, FDG-PET provides a functional view of disease-related brain changes [22].

FDG-PET is often a strong modality for AD classification, and in some studies it can perform as well as or better than multimodal combinations. This is important because multimodal learning should not be assumed to improve performance automatically [23]. A useful multimodal model must learn complementary information from MRI and PET rather than simply adding a second source of noise or overfitting to the smaller complete subset.

Despite its diagnostic value, PET has important practical limitations. It requires injection of a radioactive tracer, making it more invasive than MRI. The radiation dose limits repeated acquisitions, and the production of radiotracers requires specialized infrastructure. In addition, the short half-life of many tracers creates logistical constraints, since imaging centers must have access to nearby production facilities or reliable distribution chains. These factors make PET more expensive and less widely available than MRI. Consequently, PET is often missing in clinical datasets and even in research cohorts such as ADNI.

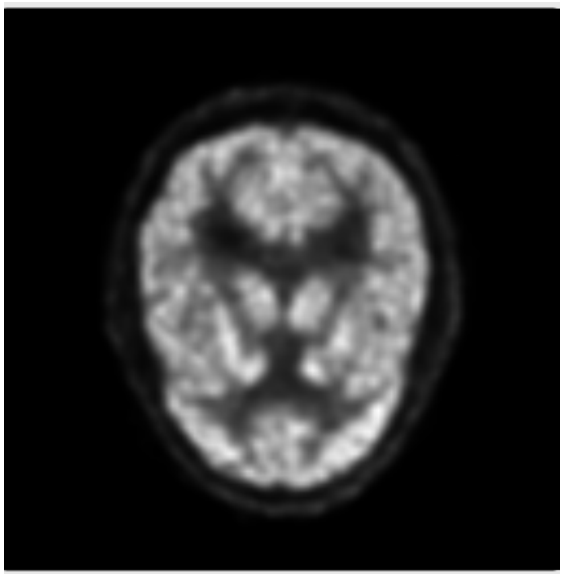


Figure 2.2: Axial slice of an FDG-PET volume from the ADNI dataset

2.4. Machine learning for Alzheimer's disease classification

Machine learning methods have been widely applied to Alzheimer's disease classification, especially using ADNI data. The most common supervised task is to classify subjects into cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD) groups. This task is useful because it provides a standardized benchmark for evaluating imaging-based models, but it should not be confused with the full clinical diagnosis of Alzheimer's disease. Clinical diagnosis is longitudinal and multimodal, involving cognitive assessments, medical history, imaging, fluid biomarkers and expert interpretation. The CN versus MCI versus AD classification task is therefore a simplified but valuable experimental setting [24].

The three-class task is also intrinsically difficult. CN and AD subjects often represent more clearly separated disease stages, while MCI is heterogeneous. Some MCI subjects remain stable for long periods, some progress to AD, and others may have cognitive impairment due to causes not primarily related to Alzheimer's pathology [25]. As a result, MCI is often the most difficult class to distinguish, and performance on the full CN versus MCI versus AD task is typically lower than on binary AD versus CN classification.

Early machine learning studies often relied on handcrafted features extracted from MRI, PET or clinical measurements. These features included regional brain volumes, cortical thickness, hippocampal measurements, voxel-based morphometry features, standardized uptake values from PET, and cognitive scores. Standard classifiers such as support vector machines, logistic regression, random forests and kernel methods were then trained on these representations. These approaches were useful because they provided interpretable pipelines and worked with relatively small sample sizes. However, their performance depended heavily on feature engineering, anatomical segmentation and preprocessing choices [26].

Deep learning changed this setting by allowing models to learn representations directly from images or minimally processed volumes. Convolutional neural networks have been widely used for MRI-based and PET-based classification, either in two-dimensional slice-based form or as three-dimensional networks operating on entire volumes [27]. These methods reduce the dependence on handcrafted features and can learn spatial patterns of atrophy or hypometabolism directly from the data. Nevertheless, deep learning in ADNI remains challenging because the number of subjects is small compared with the dimensionality of neuroimaging volumes, and because repeated visits, multiple acquisitions and class imbalance require careful evaluation protocols.

2.4.1. Multimodal learning

A large part of the multimodal deep learning literature assumes that all modalities are available for every subject. In the ADNI setting, this usually means that each training and test sample contains both MRI and FDG-PET. Under this assumption, several fusion strategies have been proposed [28].

Early approaches often used separate modality-specific feature extractors followed by a fusion layer. For example, MRI and PET features may be extracted by separate convolutional or autoencoder branches and then concatenated before classification. Other methods use staged fusion, attention mechanisms or transformer-based architectures to combine modality-specific and shared information. DiaMond [1] is an important example of this approach, recently achieving a strong performance on the three-way classification problem using vision transformers and multiple attention mechanisms. These models demonstrate that MRI and FDG-PET can be fused effectively when complete paired data are available.

However, complete-modality models have an important limitation. They are designed for the ideal case in which both modalities are present at training and inference time. This assumption is often unrealistic, especially for PET. If such a model requires MRI and FDG-PET for every subject, then subjects without PET must either be discarded, imputed, or handled by a separate model [29]. Discarding incomplete subjects reduces the amount of training data, while imputation introduces additional modeling assumptions and potential errors.

In response to this limitation, several methods have addressed incomplete multimodal learning more directly. A common strategy is to synthesize or impute the missing modality, either at the image level or in a learned feature space, so that a complete-modality classifier can still be used [29]. Other approaches learn shared latent representations, use autoencoder-based objectives or use kernel combination techniques to exploit partially observed modalities [30, 31]. These methods show that incomplete multimodal data can be used without discarding subjects, but they also highlight an additional challenge: the model must not only accept missing inputs, but also produce representations that remain compatible across different modality-availability patterns.

2.4.2. Evaluation issues in ADNI

Evaluation on ADNI requires particular care. Since ADNI is longitudinal, the same subject may appear in multiple visits and may have several scans over time. If scans from the same subject are split across training, validation and test sets, a model may indirectly recognize subject-specific anatomy rather than learning disease-related patterns. This would lead to overly optimistic estimates of generalization. For this reason, subject-level splitting is essential, meaning that all scans from the same patient should appear in only one split.

Class imbalance is another important issue. CN, MCI and AD groups are not always equally represented, and MCI is often both the largest and most heterogeneous class. Overall accuracy may therefore hide poor performance on minority or clinically important classes. Balanced accuracy is commonly used to address this issue, since it averages recall across classes and gives equal weight to each diagnostic category.

The construction of multimodal samples introduces further complications. MRI scans, PET scans and diagnostic assessments are not always acquired on the same date. Matching them requires temporal criteria, and different choices can lead to different datasets. In this thesis, the main comparisons are therefore performed between models trained and evaluated under the same preprocessing, splitting and modality-availability protocol.

2.4.3. Positioning of the present thesis

The present thesis is positioned at the intersection of multimodal learning, missing-modality learning and robust training. The goal is not only to construct a model that can technically accept missing modalities, but also to train it so that its shared classifier behaves reliably across the latent representations induced by different modality-availability patterns.

The proposed model uses a Product-of-Experts formulation to combine MRI and FDG-PET in a shared latent space. This allows the model to produce predictions from MRI alone, PET alone or from the combined modalities, using the same classifier. In this sense, the architecture addresses the missing-modality problem by design.

However, the architecture alone does not guarantee that the latent distributions induced by MRI-only and MRI-plus-PET inputs are aligned or equally suitable for classification. These different modality configurations may produce different latent-input regimes for the shared classifier. This motivates the training strategy developed later in the thesis. Inspired by domain generalization and meta-regularization, the classifier is regularized so that its parameters are encouraged to perform robustly across different modality-availability regimes. In addition, a KL-based alignment term encourages compatibility between the unimodal and multimodal latent distributions.

3

Probabilistic model

This chapter develops a probabilistic model suited to handle missing modalities at both training and inference time. Section 3.1 formally defines the multimodal classification problem, identifies the predictive distributions to be learned, and highlights the limitations of training separate classifiers for each possible combination of available modalities. Section 3.2 introduces a shared latent variable as a unified representation of the disease-relevant information, and expresses the desired predictive distributions as expectations over this unobserved variable. This separates the problem of modality fusion from the problem of classification. The section also briefly discusses the role of latent variables in generative models and clarifies how the present work instead uses them in a discriminative setting. Section 3.3 derives a Product-of-Experts formulation for the multimodal latent distribution, showing how unimodal latent distributions can be combined without introducing an independent encoder for each modality subset. Section 3.4 then specializes this framework to Gaussian densities. A first, direct Gaussian parametrization is shown to lead to a positive-definiteness constraint on the resulting covariance matrix. This issue is then resolved by adopting an MVAE-inspired reparametrization of the experts, which yields closed form unimodal and multimodal latent distributions without additional covariance constraints. Finally, Section 3.5 summarizes the resulting model, its learnable components, and the assumptions on which the construction relies.

3.1. Motivation

Standard supervised classification problems involve predicting a label $y \in \mathcal{Y}$ from a collection of observed features or modalities $x_i \in \mathcal{X}_i$, for $i = 1, \ldots, d$, by estimating the discriminative distribution $p(y|x_1, \ldots, x_d)$ using a parametrized model $\hat{p}_\theta(y|x_1, ..., x_d)$. When only a subset of modalities is observed, the relevant predictive distribution conditions only on the available variables.

In the present work, $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$ and $y \in \mathcal{Y}$ are random variables denoting the MRI scan, the FDG–PET scan and the diagnosis of a given subject, respectively, assumed to be distributed according to a true, unknown joint distribution $p(x_1, x_2, y)$. The space $\mathcal{Y} = \{0, 1, 2\}$ is the set of possible labels / diagnoses corresponding to the CN, MCI and AD disease stages, while $\mathcal{X}_1$ and $\mathcal{X}_2$ are suitable high-dimensional spaces where the MRI and FDG–PET volumes are sampled from. The aim is to learn approximations to the predictive distributions $p(y|x_i)$, for $i \in \{1, 2\}$, as well as $p(y|x_1, x_2)$ when both modalities are available.

In a hypothetical scenario where a large number of complete observations $(x_1^i, x_2^i, y^i) \overset{\text{i.i.d.}}{\sim} p(x_1, x_2, y)$ were available, one could define three separate discriminative models, namely $\hat{p}_{\theta_1}(y|x_1)$, $\hat{p}_{\theta_2}(y|x_2)$ and $\hat{p}_{\theta_3}(y|x_1, x_2)$, and train them independently to estimate the true conditionals $p(y|x_1), p(y|x_2)$ and $p(y|x_1, x_2)$. At inference time, the appropriate model could then be selected according to the available modalities.

In practice, this idealized setting is not possible for several reasons. Firstly, since FDG–PET scans are limited by cost and accessibility barriers, the majority of subjects are observed only through MRI, resulting in samples of the form (x_1^i, y^i). Secondly, even when both modalities are available for a given subject, it may happen that the acquisition dates suffer from a large temporal mismatch. Consequently, as discussed in Section 5, this may allow the disease to evolve significantly, and the scans would fail to capture a sufficiently similar brain state. For this reason, in these cases the acquisitions are treated as unpaired MRI and FDG–PET observations (x_1^i, y^i) and (x_2^i, y^i). The available dataset is therefore fragmented in complete and unpaired samples, motivating the following definition.

Definition 3.1.1 (Complete and unpaired samples sets)**.** The set of complete samples D^f is a set of N_f observations (x_1^i, x_2^i, y^i) from $p(x_1, x_2, y)$:

$$D^f := \{(x_1^i, x_2^i, y^i)\}_{i=1}^{N_f}, \qquad (x_1^i, x_2^i, y^i) \overset{\text{i.i.d.}}{\sim} p(x_1, x_2, y).$$

Further, let $p(x_1, y)$ and $p(x_2, y)$ denote the marginal distributions derived from the joint distribution $p(x_1, x_2, y)$. The set of unpaired MRI samples D^m is a set of N_m observations (x_1^i, y^i) from $p(x_1, y)$:

$$D^m := \{(x_1^i, y^i)\}_{i=1}^{N_m}, \qquad (x_1^i, y^i) \overset{\text{i.i.d.}}{\sim} p(x_1, y).$$

Similarly, the set of unpaired FDG–PET samples D^p is a set of N_p observations (x_2^i, y^i) from $p(x_2, y)$:

$$D^p := \{(x_2^i, y^i)\}_{i=1}^{N_p}, \qquad (x_2^i, y^i) \overset{\text{i.i.d.}}{\sim} p(x_2, y).$$

Remark. The MRI and FDG–PET samples from D^f, D^m, and D^p constitute distinct observations. In particular, an MRI scan appearing in a complete sample of D^f does not belong to any unpaired sample in D^m, and an FDG–PET scan appearing in a complete sample of D^f does not belong to any unpaired sample in D^p.

In the considered setting, the availability of these sample types is highly imbalanced, which we model as $N_m \gg N_f \gg N_p$. To best leverage the limited dataset, each of the models $\hat{p}_{\theta_1}(y|x_1)$, $\hat{p}_{\theta_2}(y|x_2)$ and $\hat{p}_{\theta_3}(y|x_1, x_2)$ would be trained on all the available scans for each modality combination. More specifically, the models would be trained on the corresponding modality-specific subset of samples defined as follows.

Definition 3.1.2 (Modality-available sets)**.** The set of MRI-available samples D_1 combines the MRI acquisitions from D^m and D^f:

$$D_1 := D^m \cup \{(x_1^i, y^i) : (x_1^i, x_2^i, y^i) \in D^f\}.$$

The set of FDG–PET-available samples D_2 combines the FDG–PET acquisitions from D^p and D^f:

$$D_2 := D^p \cup \{(x_2^i, y^i) : (x_1^i, x_2^i, y^i) \in D^f\}.$$

Finally, the set of complete multimodal samples is

$$D_{1,2} := D^f.$$

Therefore, $\hat{p}_{\theta_1}(y|x_1)$ would be trained on D_1, $\hat{p}_{\theta_2}(y|x_2)$ on D_2, and $\hat{p}_{\theta_3}(y|x_1, x_2)$ on $D_{1,2}$. Importantly, however, the scarce availability of complete samples and FDG–PET-available samples poses significant challenges for this separate-model strategy. The multimodal classifier would be estimated exclusively from the small complete subset $D_{1,2}$, comprising only N_f samples, making it prone to overfitting and unreliable generalization. A similar issue would affect the FDG-PET classifier, having only $N_f + N_p$ available samples from D_2. Moreover, the abundant unpaired MRI samples in D^m would remain unused for learning the FDG-PET and multimodal predictors, even though they contain information about the same underlying pathology: structural brain changes visible on MRI are correlated with the metabolic changes captured by FDG–PET, and

both are driven by the disease process. Separate models also create a practical clinical dilemma. For a patient with both scans, the individual MRI and FDG-PET classifiers may assign different diagnostic probabilities, while the modeling framework itself provides no probabilistic mechanism for reconciling these predictions.

These limitations motivate a unified framework in which the unimodal and multimodal predictive distributions are coupled through shared parameters and a common representation. A principled way to construct such a framework is to introduce an unobserved shared variable capturing disease-relevant information common to the observed modalities.

3.2. Shared latent representation

To construct a unified framework, we introduce a latent random variable $\mathbf{z} \in \mathbb{R}^L$, for which no direct observations are available, along with the core assumption that MRI, FDG–PET and the diagnosis are conditionally independent given it. At this stage, $\mathbf{z}$ should be understood as a modeling device that defines a common latent space in which information from different modalities can be expressed and combined.

The key advantage of this modelling choice, together with the conditional independence assumption, is that it allows us to decouple classification from modality fusion. In fact, the latent variable acts as an intermediate pathway between the modalities and the diagnosis, so that $p(y|x_i)$ and $p(y|x_1, x_2)$ can be broken down in two groups of terms: $p(\mathbf{z}|x_i)$ and $p(\mathbf{z}|x_1, x_2)$ which relate x_1, x_2 to $\mathbf{z}$, and $p(y|\mathbf{z})$ which instead relates $\mathbf{z}$ to y. Therefore, while $p(y|\mathbf{z})$ acts as a classifier by predicting the diagnosis y given the latent $\mathbf{z}$, the role of the modality-specific components $p(\mathbf{z}|x_i)$ and $p(\mathbf{z}|x_1, x_2)$ is to construct a distribution over $\mathbf{z}$ from the available observations. If only MRI is available, this distribution is based on x_1; if only FDG–PET is available, it is based on x_2; and if both modalities are available, the information from both scans is fused into a single latent distribution. In this way, all the predictive distributions share the same final classifier, while differing only in how the latent representation is obtained.

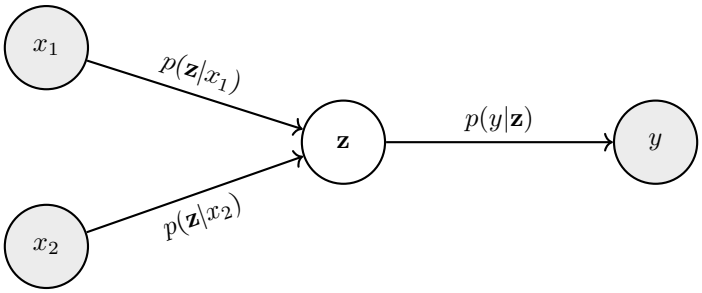


Figure 3.1: Latent-variable view of the proposed probabilistic model. The observed modalities x_1 and x_2 provide information about the shared latent representation $\mathbf{z}$, which is then used to predict the diagnosis y.

Formally, the conditional independence assumption can be formulated as follows:

$$p(x_1, x_2, y|\mathbf{z}) = p(x_1|\mathbf{z})p(x_2|\mathbf{z})p(y|\mathbf{z}).$$

Under this assumption, it is possible to express the predictive distributions $p(y|x_i)$ and $p(y|x_1, x_2)$ in terms of $p(y|\mathbf{z})$, as shown in the following proposition.

Proposition 3.2.1 (Predictive distributions under the latent representation)**.** Under the conditional independence assumption, the predictive distributions may be expressed as

$$p(y|x_i) = \int p(y|\mathbf{z})p(\mathbf{z}|x_i)d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|x_i)}[p(y|\mathbf{z})] \quad i \in \{1, 2\}$$

$$p(y|x_1, x_2) = \int p(y|\mathbf{z})p(\mathbf{z}|x_1, x_2)d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|x_1,x_2)}[p(y|\mathbf{z})]$$

Proof. We begin from the case of $p(y|x_i)$ for $i \in \{1, 2\}$.

$$p(y, \mathbf{z}|x_i) = \frac{p(x_i, y|\mathbf{z})p(\mathbf{z})}{p(x_i)} = p(y|\mathbf{z})\frac{p(x_i|\mathbf{z})p(\mathbf{z})}{p(x_i)} = p(y|\mathbf{z})p(\mathbf{z}|x_i)$$

$$p(y|x_i) = \int p(y, \mathbf{z}|x_i)d\mathbf{z} = \int p(y|\mathbf{z})p(\mathbf{z}|x_i)d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|x_i)}[p(y|\mathbf{z})]$$

An analogous approach can be used to obtain a similar expression for $p(y|x_1, x_2)$:

$$p(y|x_1, x_2) = \int p(y, \mathbf{z}|x_1, x_2)d\mathbf{z} = \int p(y|\mathbf{z})p(\mathbf{z}|x_1, x_2)d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|x_1,x_2)}[p(y|\mathbf{z})]$$

□

Notably, these expressions reflect the two-stage relation between y and x_1, x_2 mediated by $\mathbf{z}$ described earlier. First, the distribution of $\mathbf{z}$ is obtained by means of either $p(\mathbf{z}|x_i)$ or $p(\mathbf{z}|x_1, x_2)$, according to the availability of the modalities. Secondly, the probability distribution on y is obtained by taking the expectation of $p(y|\mathbf{z})$ with respect to $\mathbf{z}$. This double stage inference process is illustrated in the graph in Figure 3.1.

The latent variable $\mathbf{z}$ may be interpreted as a representation of the patient's unobserved brain state, including structural, biological and neurodegenerative factors relevant for diagnosis. Under this interpretation, MRI and FDG–PET are modality-specific and partial views of the same underlying state. The conditional independence assumption then expresses the idea that, once $\mathbf{z}$ is known, the individual modalities and the diagnosis provide no additional information about each other. Indeed, for $i, j \in \{1, 2\}$ with $i \neq j$, this for instance implies that

$$p(x_i|x_j, \mathbf{z}) = p(x_i|\mathbf{z}), \qquad p(x_i|y, \mathbf{z}) = p(x_i|\mathbf{z}), \qquad p(y|x_i, \mathbf{z}) = p(y|\mathbf{z}).$$

This assumption is idealized and should not be interpreted as an exact description of the data-generating process. For instance, MRI and FDG–PET may share site-specific noise, scanner effects, preprocessing artefacts or other confounding factors, which can induce dependence even after conditioning on disease state. Nevertheless, conditional independence is a common and useful modelling assumption in multimodal latent-variable models, including MVAEs [32]. Moreover, as shown in the next section, it leads to a natural formulation for combining the unimodal conditional distributions $p(\mathbf{z}|x_1)$ and $p(\mathbf{z}|x_2)$ into a multimodal distribution $p(\mathbf{z}|x_1, x_2)$, which, as we will see, allows us to avoid learning a parametrized latent distribution for every possible modality subset.

3.3. Product of Experts

The expressions derived above show that all predictive distributions depend on the same conditional distribution $p(y|\mathbf{z})$, but differ in the latent distribution with respect to which this conditional is averaged. It therefore remains to characterize the multimodal latent conditional $p(\mathbf{z}|x_1, x_2)$ in relation to the unimodal conditionals $p(\mathbf{z}|x_1)$ and $p(\mathbf{z}|x_2)$.

A possible approach would be to treat $p(\mathbf{z}|x_1, x_2)$ as an independent object, unrelated to the unimodal latent conditionals. However, this would partly reintroduce the limitations of the separate-model strategy. The multimodal pathway would then require its own latent distribution, learned only through complete MRI–PET observations, and its relation to the unimodal pathways would have to be imposed by design rather than following from the probabilistic assumptions. The introduction of the latent variable would merely shift the original problem from approximating separate predictive distributions $p(y|x_i)$ and $p(y|x_1, x_2)$ to learning separate approximations to latent conditionals $p(\mathbf{z}|x_i)$ and $p(\mathbf{z}|x_1, x_2)$.

Instead, the conditional independence assumption provides a direct relationship between these quantities. Under this assumption, the multimodal latent conditional can be expressed in terms

of the unimodal latent conditionals and the prior over $\mathbf{z}$. Specifically, as shown below, $p(\mathbf{z}|x_1, x_2)$ is proportional to the product of $p(\mathbf{z}|x_1)$ and $p(\mathbf{z}|x_2)$, corrected by the prior $p(\mathbf{z})$. This identity is derived at the level of the ideal, unknown distributions. Later, since these conditionals are not available in practice, the same structure will be used as a blueprint for defining the learned multimodal approximation from learned unimodal encoder distributions.

In the remainder of this section, we present an adaptation to our specific case of the more general framework derived in [32].

Proposition 3.3.1 (Product-of-Experts latent conditional)**.** Under the conditional independence assumption, the multimodal latent conditional satisfies

$$p(\mathbf{z}|x_1, x_2) \propto \frac{p(\mathbf{z}|x_1)p(\mathbf{z}|x_2)}{p(\mathbf{z})}. \tag{3.1}$$

Proof. Rewriting $p(\mathbf{z}|x_1, x_2)$ using the definition of conditional probability, Bayes' rule, and conditional independence gives

$$p(\mathbf{z}|x_1, x_2) = \frac{p(x_1, x_2|\mathbf{z})p(\mathbf{z})}{p(x_1, x_2)} \frac{p(\mathbf{z})}{p(x_1, x_2)} p(x_1|\mathbf{z})p(x_2|\mathbf{z}).$$

Using Bayes' rule again,

$$p(x_1|\mathbf{z}) = \frac{p(\mathbf{z}|x_1)p(x_1)}{p(\mathbf{z})}, \qquad p(x_2|\mathbf{z}) = \frac{p(\mathbf{z}|x_2)p(x_2)}{p(\mathbf{z})}.$$

Substituting these expressions yields

$$p(\mathbf{z}|x_1, x_2) = \frac{p(x_1)p(x_2)}{p(x_1, x_2)} \frac{p(\mathbf{z}|x_1)p(\mathbf{z}|x_2)}{p(\mathbf{z})}.$$

Since the factor

$$C(x_1, x_2) = \frac{p(x_1)p(x_2)}{p(x_1, x_2)}$$

does not depend on $\mathbf{z}$, we obtain

$$p(\mathbf{z}|x_1, x_2) \propto \frac{p(\mathbf{z}|x_1)p(\mathbf{z}|x_2)}{p(\mathbf{z})}.$$

□

The omitted constant of proportionality does not depend on $\mathbf{z}$ and ensures the distribution integrates to one. This last formula is known as *Product of Experts* [32], as it combines the beliefs on the latent variable of two unimodal experts $p(\mathbf{z}|x_1)$ and $p(\mathbf{z}|x_2)$ by means of a product, corrected by the prior $p(\mathbf{z})$.

This formulation gives a natural mechanism for modality fusion. If only modality $i \in \{1, 2\}$ is available, the corresponding unimodal latent distribution is used. If both modalities are available, the two expert beliefs are combined through Equation (3.1). Thus, the latent variable provides a common space for prediction, and the Product of Experts provides a probabilistic rule for fusing modality-specific information within that space.

It is important, however, to clarify the role of this derivation in the present work. Latent variable models are often introduced in a generative setting, where one explicitly parameterizes the generative likelihoods of the observed variables given the latent variable

$$r_\eta(x_1, x_2, y, \mathbf{z}) = r_\eta(x_1, x_2, y|\mathbf{z})r(\mathbf{z}).$$

Training would then typically involve maximizing the marginal likelihood

$$r_\eta(x_1, x_2, y) = \int r_\eta(x_1, x_2, y|\mathbf{z})r(\mathbf{z})d\mathbf{z},$$

or a variational lower bound on it, using an approximate posterior such as $q_\psi(\mathbf{z}|x_1, x_2, y)$ to deal with the intractability of the true posterior.

However, this is not the approach taken here. Our goal is not to model the distribution of MRI or FDG–PET scans themselves, but to estimate the discriminative predictive distributions $p(y|x_i)$ and $p(y|x_1, x_2)$. Explicitly modeling $p(x_i|\mathbf{z})$ would require high-dimensional image decoders and an additional variational-inference problem, while these generative components would only be used indirectly for the final classification task. For this reason, we instead use the latent variable assumptions as a modeling principle for constructing a discriminative classifier.

In particular, the unknown true distributions $p(\mathbf{z}|x_i)$ and $p(y|\mathbf{z})$ are approximated by the learned distributions $\hat{p}_{\phi_i}(\mathbf{z}|x_i)$ and $\hat{p}_\theta(y|\mathbf{z})$, all parametrized by neural networks. The multimodal latent distribution is not parameterized independently; rather, it is defined by applying the Product-of-Experts structure from Equation (3.1) to the learned unimodal approximations. Letting $\phi = \{\phi_1, \phi_2\}$, we define

$$\hat{p}_\phi(\mathbf{z}|x_1, x_2) := \frac{1}{Z(x_1, x_2)} \frac{\hat{p}_{\phi_1}(\mathbf{z}|x_1)\hat{p}_{\phi_2}(\mathbf{z}|x_2)}{p(\mathbf{z})}, \tag{3.2}$$

where $Z(x_1, x_2)$ is the normalizing constant. The prior $p(\mathbf{z})$ is specified rather than learned. Thus, the Product-of-Experts identity is not assumed to give direct access to the true multimodal conditional; instead, it motivates the architecture used to approximate it. The learned predictive distributions are then defined by mirroring the ideal expectation identities derived earlier:

$$\hat{p}_{\theta,\phi_i}(y|x_i) := \mathbb{E}_{\hat{p}_{\phi_i}(\mathbf{z}|x_i)}\left[\hat{p}_\theta(y|\mathbf{z})\right], \qquad i \in 1, 2,$$

and

$$\hat{p}_{\theta,\phi}(y|x_1, x_2) := \mathbb{E}_{\hat{p}_\phi(\mathbf{z}|x_1,x_2)}\left[\hat{p}_\theta(y|\mathbf{z})\right].$$

These equations define the learned discriminative model. They should therefore be understood as approximations to the ideal predictive distributions $p(y|x_i)$ and $p(y|x_1, x_2)$, not as exact equalities involving the true population distribution.

In this way, the proposed formulation allows all available observations to contribute to the learned model according to the modalities they contain. Samples with only MRI update the MRI encoder $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$ and the shared classifier $\hat{p}_\theta(y|\mathbf{z})$. Samples with only FDG–PET analogously update the FDG–PET encoder $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$ and the same shared classifier. Complete samples contribute to both unimodal pathways and additionally train the multimodal prediction pathway through the Product-of-Experts distribution $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$. Hence, incomplete and complete observations can be exploited within a single framework, without requiring a separate classifier for each modality-availability pattern.

The formulation is therefore symmetric at the level of the model: MRI, FDG–PET and MRI–FDG–PET inputs are all handled through the same latent-space classifier and differ only in the latent distribution used to represent the available information. The asymmetry remains at the level of the data, since MRI observations are substantially more frequent than FDG–PET and complete observations. Consequently, the different pathways may receive different amounts of supervision during training. This imbalance motivates the training strategies introduced in the next chapter, which aim to improve robustness across the latent-input regimes induced by different modality subsets.

3.4. Parametrization

3.4.1. Gaussian parametrization

For arbitrary choices of $\hat{p}_{\phi_i}(\mathbf{z}|x_i)$ the normalization constant in Equation 3.2 is generally not available in closed-form. To obtain a tractable model, we assume that the prior $p(\mathbf{z})$ is normally

distributed, and restrict the choice of densities for $\hat{p}_{\phi_i}(\mathbf{z}|x_i)$ to the Gaussian family. This choice is convenient because products (and quotients) of Gaussian densities are proportional to Gaussian densities, so the resulting distribution can be computed analytically. As a first attempt, suppose that the prior and the unimodal conditional distributions are given by

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I), \qquad \hat{p}_{\phi_i}(\mathbf{z}|x_i) = \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i), \qquad i \in \{1, 2\}.$$

Recall that if $p_1(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$ and $p_2(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$, then their product is proportional to a Gaussian density [33],

$$p_1(x)p_2(x) \propto \mathcal{N}(x; \mu, \Sigma),$$

with

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}, \qquad \mu = \Sigma(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2).$$

In the simplified isotropic case, where $\Sigma_1 = \sigma_1^2 I$ and $\Sigma_2 = \sigma_2^2 I$, it is easy to see that the resulting covariance and mean are

$$\sigma^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}, \qquad \mu = \sigma^2 \left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right).$$

The mean of the product is therefore a precision-weighted average of the expert means. An expert with smaller variance has larger precision and contributes more strongly to the fused mean. This provides an intuitive interpretation of the Product of Experts: more confident experts have greater influence on the resulting latent belief.

Returning to Equation (3.2), the division by the prior introduces a negative contribution in precision space. With the direct Gaussian parametrization above, the multimodal density $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$ is proportional to a Gaussian with covariance and mean

$$\Sigma_{\text{PoE}} = (\Sigma_1^{-1} + \Sigma_2^{-1} - I)^{-1}, \qquad \mu_{\text{PoE}} = \Sigma_{\text{PoE}}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2).$$

This formulation is however mathematically inconvenient. For Σ_{PoE} to define a valid covariance matrix, the precision matrix $\Sigma_1^{-1} + \Sigma_2^{-1} - I$ must be symmetric positive definite. This condition is not automatically satisfied by arbitrary valid covariance matrices Σ_1 and Σ_2. In the isotropic case, it reduces to

$$\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - 1 > 0.$$

Thus, the neural networks would have to output covariance matrices satisfying an additional constraint. Enforcing such a constraint during training is inconvenient and may lead to numerical instability. This motivates a different Gaussian parametrization.

3.4.2. MVAE-Q reparametrization

To avoid the positive-definiteness constraint arising from the direct Gaussian parametrization, we follow the reparametrization used in the MVAE framework [32]. Instead of letting the neural network encoder directly output the parameters of the unimodal conditional distribution $\hat{p}_{\phi_i}(\mathbf{z}|x_i)$, each encoder outputs the parameters of an auxiliary Gaussian expert

$$\tilde{p}_{\phi_i}(\mathbf{z}|x_i) = \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i).$$

The actual unimodal conditional distribution is then defined by combining this auxiliary expert with the prior:

$$\hat{p}_{\phi_i}(\mathbf{z}|x_i) \propto \tilde{p}_{\phi_i}(\mathbf{z}|x_i)p(\mathbf{z}). \tag{3.3}$$

Since both factors are Gaussian densities, the unimodal conditional distribution is Gaussian. With $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$, its covariance and mean are

$$\Sigma_i^{\text{post}} = (\Sigma_i^{-1} + I)^{-1}, \qquad \mu_i^{\text{post}} = \Sigma_i^{\text{post}}\Sigma_i^{-1}\mu_i.$$

Substituting Equation (3.3) into Equation (3.2) gives

$$\hat{p}_\phi(\mathbf{z}|x_1, x_2) \propto \frac{\tilde{p}_{\phi_1}(\mathbf{z}|x_1)p(\mathbf{z})\tilde{p}_{\phi_2}(\mathbf{z}|x_2)p(\mathbf{z})}{p(\mathbf{z})}.$$

Therefore,

$$\hat{p}_\phi(\mathbf{z}|x_1, x_2) \propto \tilde{p}_{\phi_1}(\mathbf{z}|x_1)\tilde{p}_{\phi_2}(\mathbf{z}|x_2)p(\mathbf{z}).$$

The prior now contributes positively in precision space rather than negatively. Consequently, the multimodal conditional distribution is Gaussian with covariance and mean

$$\Sigma_{\text{PoE}} = (\Sigma_1^{-1} + \Sigma_2^{-1} + I)^{-1}, \qquad \mu_{\text{PoE}} = \Sigma_{\text{PoE}}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2). \tag{3.4}$$

As long as Σ_1 and Σ_2 are valid covariance matrices, the matrix $\Sigma_1^{-1} + \Sigma_2^{-1} + I$ is symmetric positive definite. Hence, the covariance matrix in Equation (3.4) is always well defined, and no additional constraints are required beyond ensuring that the auxiliary covariance matrices are valid. This construction is formalized in the following definition, and its closed-form Gaussian parameters are given in Proposition 3.4.1.

Definition 3.4.1 (MVAE-Q Product-of-Experts parametrization)**.** Let $S \subseteq \{1, 2\}$ denote a non-empty set of available modalities, and let $x_S = \{x_i : i \in S\}$. For each modality $i \in S$, let the encoder define an auxiliary Gaussian expert

$$\tilde{p}_{\phi_i}(\mathbf{z}|x_i) = \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i),$$

and let the prior be $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$. The latent conditional distribution associated with the available modality set S is defined as

$$\hat{p}_\phi(\mathbf{z}|x_S) \propto p(\mathbf{z}) \prod_{i \in S} \tilde{p}_{\phi_i}(\mathbf{z}|x_i).$$

Proposition 3.4.1 (Closed-form Gaussian latent conditional)**.** Under the parametrization above, the latent conditional distribution $\hat{p}_\phi(\mathbf{z}|x_S)$ is Gaussian:

$$\hat{p}_\phi(\mathbf{z}|x_S) = \mathcal{N}(\mathbf{z}; \mu_S, \Sigma_S),$$

where

$$\Sigma_S = \left(I + \sum_{i \in S} \Sigma_i^{-1}\right)^{-1}, \qquad \mu_S = \Sigma_S \left(\sum_{i \in S} \Sigma_i^{-1}\mu_i\right).$$

Proof. The result follows from the fact that the product of Gaussian densities is proportional to a Gaussian density. Since the prior contributes precision matrix I and each auxiliary expert contributes precision matrix Σ_i^{-1}, the resulting precision matrix is

$$I + \sum_{i \in S} \Sigma_i^{-1}.$$

The corresponding covariance is therefore

$$\Sigma_S = \left(I + \sum_{i \in S} \Sigma_i^{-1} \right)^{-1}.$$

The natural parameter of the product is given by the sum of the expert natural parameters,

$$\sum_{i \in S} \Sigma_i^{-1} \mu_i,$$

since the prior has zero mean and therefore contributes no linear term. Hence,

$$\mu_S = \Sigma_S \left(\sum_{i \in S} \Sigma_i^{-1} \mu_i \right).$$

□

For $S = \{i\}$, this recovers the reparametrized unimodal conditional distribution in Equation (3.3). For $S = \{1, 2\}$, it recovers the multimodal Product-of-Experts distribution in Equation (3.4). This notation also makes clear how the framework would extend to more than two modalities.

3.5. Final model

The final model can now be summarized in a compact form. For any available modality set $S \subseteq \{1, 2\}$, the model first constructs a latent conditional distribution

$$\hat{p}_\phi(\mathbf{z}|x_S) \propto p(\mathbf{z}) \prod_{i \in S} \tilde{p}_{\phi_i}(\mathbf{z}|x_i),$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$ is the prior and each auxiliary expert $\tilde{p}_{\phi_i}(\mathbf{z}|x_i) = \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i)$ is produced by a modality-specific encoder. The diagnostic predictive distribution is then defined as

$$\hat{p}_{\theta,\phi}(y|x_S) = \mathbb{E}_{\hat{p}_\phi(\mathbf{z}|x_S)} \left[\hat{p}_\theta(y|\mathbf{z}) \right]. \tag{3.5}$$

The learnable components of the model are:

- an MRI encoder, parameterized by ϕ_1, which maps x_1 to the parameters μ_1 and Σ_1 of the auxiliary Gaussian expert $\tilde{p}_{\phi_1}(\mathbf{z}|x_1)$;
- an FDG–PET encoder, parameterized by ϕ_2, which maps x_2 to the parameters μ_2 and Σ_2 of the auxiliary Gaussian expert $\tilde{p}_{\phi_2}(\mathbf{z}|x_2)$;
- a shared classifier, parameterized by θ, which maps latent representations $\mathbf{z}$ to diagnostic probabilities through $\hat{p}_\theta(y|\mathbf{z})$.

In practice, the covariance matrices are taken to be diagonal, so each encoder outputs a mean vector and a vector of marginal variances or log-variances. This reduces the number of parameters and keeps the Gaussian computations tractable in the latent space.

The model provides a unified prediction mechanism under missing modalities. If only MRI is available, then $S = \{1\}$ and the prediction uses $\hat{p}_\phi(\mathbf{z}|x_1)$. If only FDG–PET is available, then $S = \{2\}$ and the prediction uses $\hat{p}_\phi(\mathbf{z}|x_2)$. If both modalities are available, then $S = \{1, 2\}$ and the prediction uses the Product-of-Experts distribution $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$. In all cases, the same classifier $\hat{p}_\theta(y|\mathbf{z})$ is used.

The formulation relies on several assumptions. First, the latent variable $\mathbf{z}$ is assumed to capture disease-relevant information shared across MRI, FDG–PET and diagnosis. Second, the conditional independence assumption is used to motivate the Product-of-Experts fusion rule, even though it may be violated in practice by site effects, modality-specific noise or other confounding factors.

Third, the Gaussian parametrization imposes a specific form on the latent uncertainty, which may not fully capture the true structure of uncertainty in the data. Finally, the Product-of-Experts architecture defines how modality-specific information is fused, but it does not by itself guarantee that the latent distributions induced by different modality subsets are perfectly aligned. This last point also motivates the training strategy developed in the next chapter.

4

Meta-learning training

This chapter develops the training strategy for the Product-of-Experts model. Section 4.1 motivates the need for it, arguing that different modality-availability patterns induce different latent-input regimes for the shared classifier. Section 4.2 derives a meta-learned regularizer for the classifier parameters, designed to favour updates that transfer across these regimes, while Section 4.3 introduces a Kullback-Leibler term encouraging the unimodal and multimodal latent distributions to remain compatible. Section 4.4 presents the resulting training algorithms.

4.1. Motivation

The probabilistic framework developed in the previous chapter couples the predictive distributions $p(y|x_1, x_2)$ and $p(y|x_i)$ by means of a shared set of parameters, solving the architectural problem of defining a unique model able to handle modality-complete and modality-incomplete samples. While this approach unifies part of the parameters, leverages the entire dataset and avoids independent models, it also introduces important challenges during the training process.

As a consequence of the two-stage decomposition of the prediction process, which separates modality fusion from classification, different modality-availability patterns can induce multiple feature domains inside the model. Specifically, introducing the latent variable $\mathbf{z}$ allows the definition of a single classifier $\hat{p}_\theta(y|\mathbf{z})$ that is shared across all prediction settings. The input to this classifier, however, may be sampled from different conditional latent distributions, namely $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$, $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$ or $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$, depending on which modalities are observed.

Consequently, each modality-availability pattern can induce a different distribution in latent space. If only MRI is available, prediction can only be performed through the MRI-induced latent distribution $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$. Similarly, if only FDG-PET is available, the latent variable is obtained through $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$. Finally, for complete samples, prediction is instead performed through the multimodal Product-of-Experts distribution $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$. Although these latent distributions are intended to describe the same underlying state, they are produced by different experts or by different combinations of experts, and may therefore differ systematically. This creates a form of domain shift of $\mathbf{z}$ across modality configurations, to which even deep neural networks such as $\hat{p}_\theta(y|\mathbf{z})$ are known to be sensitive [34]. In this sense, the missing-modality problem is not only an architectural problem, addressed by the Product-of-Experts formulation, but also a training problem: the shared classifier should learn decision rules that are stable across the latent distributions induced by different modality subsets.

These latent distributions can be interpreted as different input domains for the shared classifier $\hat{p}_\theta(y|\mathbf{z})$. The model should ideally achieve comparable performance on all of these regimes, rather than relying excessively on one of them. Whether this happens in practice depends strongly on the training procedure, since such cross-regime stability is generally not enforced by the model architecture alone.

A standard empirical loss minimization training strategy would be to apply a loss criterion, such as cross-entropy, to each predictive distribution and to combine the resulting terms in a weighted sum. Concretely, for each non-empty modality subset $S \subseteq \{1, 2\}$, let

$$\mathcal{L}_S(\theta, \phi) := \sum_i \mathcal{L}_{\text{CE}} \left(\hat{p}_{\theta,\phi}(\cdot|x_S^i), y^i \right),$$

where the sum is taken over the samples for which the modalities in S are available. More specifically, these are the modality-specific training sets defined in Definition 3.1.2: D_1 contains all the available MRI samples, D_2 all the FDG-PET samples, and $D_{1,2}$ contains the paired multimodal samples. A conventional weighted empirical loss minimization objective would then be written as

$$\min_{\theta,\phi} \quad \lambda_1 \mathcal{L}(\phi, \theta; D_1) + \lambda_2 \mathcal{L}(\phi, \theta; D_2) + \lambda_{1,2} \mathcal{L}(\phi, \theta; D_{1,2})$$

Although this objective trains under all combinations of available modalities, it does not explicitly encourage the classifier to behave consistently across the corresponding latent domains induced by the modality-specific sets D_1, D_2 and $D_{1,2}$. Informally, these domains correspond to latent variables sampled as

$$Z_1 \sim \hat{p}_{\phi_1}(\mathbf{z}|x_1) \qquad Z_2 \sim \hat{p}_{\phi_2}(\mathbf{z}|x_2) \qquad Z_{1,2} \sim \hat{p}_{\phi}(\mathbf{z}|x_1, x_2)$$

where the inputs x_i are drawn from the corresponding subset D_1, D_2 or $D_{1,2}$. In particular, the gradients induced by the different loss terms may not be aligned, and the optimization process may proceed asymmetrically by reducing preferentially the loss that is easier to minimize [2]. As a result, the model may fail to exploit common disease-related signals, may generalize unevenly across the latent domains and may yield suboptimal performance under some modality-availability regimes.

A second, closely related issue concerns the structure of the latent representation itself. Because MRI and FDG-PET contain different and only partially overlapping information, the corresponding latent conditionals should not be forced to be identical in all respects. On the other hand, it is important that such distributions agree on the representation of common information available in both modalities. For instance, AD-relevant signals should share a common representation across different distributions, so that when the two Gaussian experts agree, the intersection of the representations is preserved and the signals are properly conveyed to the shared classifier. Therefore, encouraging compatibility between the multiple conditional distributions of $\mathbf{z}$ may help to capture common patterns and correlations between modalities.

Additionally, because the goal is to achieve best possible performance across all available modality patterns, encouraging the alignment of the latent distributions may help the MRI-specific representation to preserve disease-relevant structure that is also present in the multimodal one. In this sense, obtaining a uniform representation is also closely related to the field of knowledge distillation, where embeddings produced by a teacher network with superior information should resemble as much as possible those of a student network, operating instead in a limited-information setting [35].

To address these issues, we propose a two-sided regularization strategy. The first regularization term acts on the parameters of the shared classifier $\hat{p}_\theta(y|\mathbf{z})$ and is designed to reduce overfitting to a single latent-input regime. Rather than manually choosing a fixed regularizer, we parameterize a regularizing function $R_\psi(\theta)$ and learn its parameters through a meta-learning procedure for domain generalization. In this procedure, the modality-availability regimes are used as meta-train and meta-test domains, so that the learned regularizer encourages classifier updates that transfer across regimes.

The second regularization term acts directly on the latent distributions. It uses a Kullback–Leibler divergence term to encourage compatibility between the MRI-specific latent distribution $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$, the PET-specific distribution $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$, and the multimodal distribution $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$. This alignment is applied only as a soft constraint: it encourages the distributions to agree on shared disease-relevant information while still allowing the multimodal distribution to encode additional information provided by the combination of MRI and FDG-PET.

The resulting training procedure therefore has two complementary objectives. First, it aims to learn a shared classifier that generalizes well across the latent domains induced by different modality patterns. Second, it aims to promote a coherent latent representation in which disease-related information is encoded compatibly across modalities. After the regularizer $R_\psi(\theta)$ has been meta-learned, the full Product-of-Experts model is trained from scratch using a regularized objective that combines supervised classification losses, the frozen learned classifier regularizer, and KL-based latent-distribution alignment terms.

4.2. Meta-learning regularization for cross-regime generalization

In the following section we describe the derivation of a regularizer function for the parameters of the shared classifier $\hat{p}_\theta(y|\mathbf{z})$. As motivated above, the goal is to avoid over-reliance on a single latent-input regime and to encourage comparable performance on the latent domains Z_1, Z_2 and $Z_{1,2}$. We first introduce a meta-learning training paradigm from the field of domain generalization, which provides an alternative to conventional single-domain regularization when multiple data domains are present. We then show how this idea can be combined with regularization in a data-driven approach, leading to a learned regularizer R_ψ tailored to the latent domain setting of the proposed Product-of-Experts model.

4.2.1. Meta-learning for domain generalization

Regularization is a conventional technique used to prevent overfitting of parametric models and to improve performance on unseen test samples. A large number of regularization approaches have been studied for deep neural networks, including weight penalties, dropout-based methods and data augmentation strategies. However, most of these techniques are designed for the single-domain setting, where training and test samples are assumed to share a common underlying distribution [3]. Therefore, in the present work, conventional regularization techniques alone may be insufficient, since they do not explicitly account for the fact that $\mathbf{z}$ is produced by different conditional distributions according to the available modalities.

Generalizing across domains is instead the primary goal of domain generalization. In this setting, a model is trained on a set of source domains $\mathcal{S}$ and should generalize to test domains $\mathcal{T}$, without adapting its parameters at test time. Although our setting does not correspond exactly to a classical domain generalization problem with unseen external target domains, the same idea is useful: the model should learn parameters that perform well across the latent-input domains Z_1, Z_2 and $Z_{1,2}$ induced by different modality configurations.

A possible approach to this challenge is to employ a meta-learning training paradigm [2, 36]. In particular, to mimic the distribution shift across $\mathcal{S}$ and $\mathcal{T}$, the source domains are further subdivided into *meta-train* and *meta-test* domains denoted respectively by $\bar{S}$ and $\check{S}$. A specific training objective is then introduced with the goal of minimizing losses on meta-train and meta-test domains together, in a more coordinated way.

This idea is formalized in Meta-Learning for Domain Generalization (MLDG) [2]. More specifically, this is achieved in two steps: first, the model parametrized by Θ is trained with a single SGD step on the meta-train loss $\mathcal{F}(\Theta) = \mathcal{F}(\Theta; \bar{S})$, yielding the adapted parameters $\Theta^* = \Theta - \alpha\nabla_\Theta\mathcal{F}$. After this, in the meta-testing stage, the model with the updated weights is tested on the meta-test domains $\check{S}$, producing the loss $\mathcal{G}(\Theta^*) = \mathcal{G}(\Theta^*; \check{S})$. The minimization objective is then defined to be a linear combination of these losses:

$$\mathcal{F}(\Theta) + \beta\mathcal{G}(\Theta - \alpha\nabla_\Theta\mathcal{F})$$

Intuitively, while optimizing losses of both domains altogether, this two-steps process encourages the optimizer to exploit gradient directions that, while minimizing the meta-train loss $\mathcal{F}$, are also beneficial in minimizing the meta-test loss $\mathcal{G}$. This is further motivated by analyzing a first order

Taylor expansion of $\mathcal{G}$ centered at $x = \Theta$ and evaluated at $\dot{x} = \Theta - \alpha\nabla_\Theta\mathcal{F}(\Theta)$:

$$\mathcal{G}(\Theta - \alpha\nabla_\Theta\mathcal{F}) \approx \mathcal{G}(\Theta) - \alpha\nabla_\Theta\mathcal{G}(\Theta) \cdot \nabla_\Theta\mathcal{F}(\Theta)$$

This highlights the fact that the optimizer is encouraged to find a route to a minimum where the gradients $\nabla_\Theta\mathcal{F}$ and $\nabla_\Theta\mathcal{G}$ point in compatible directions (as their dot product is encouraged to be large), thus reducing overfitting to a single loss. In the context of the present work, this is precisely the type of behaviour desired for the shared classifier: updates that improve one latent-input regime should not degrade, and ideally should also improve, the other regime.

We do not directly apply MLDG as the final training objective. Rather, we use its principle to motivate the learning of a regularizer. The role of this regularizer is to bias the classifier toward parameter regions where optimization on one modality-specific regime transfers well to another.

4.2.2. Meta-regularization

Inspired by the previous idea and by MetaReg [3], we now construct a regularizer function that is itself learned through a meta-learning procedure. The purpose is to obtain a regularizer that is not manually designed, but adapted to the cross-regime structure of our setting.

In a standard empirical risk minimization scenario, a regularizer is added to the supervised loss of the model. If M_θ denotes a classifier with parameters θ, the regularized loss can be written as

$$L_{\text{reg}}(\theta) = L(\theta) + R(\theta)$$

where $L(\theta)$ is the task loss, for instance cross-entropy, and $R(\theta)$ is a predefined penalty such as an L^1 or L^2 norm. However, designing a regularizer that is effective across domains is difficult in practice, and its usefulness may depend on the specific domain structure of the data.

This issue is addressed by replacing the fixed regularizer $R(\theta)$ with a parameterized regularizer $R_\psi(\theta)$, whose parameters are learned specifically to improve cross-domain generalization [3]. The core idea is to create a meta-learning loop in which a meta-train domain $\bar{S}$ and a meta-test domain $\check{S}$ are selected from the available source domains. The model is meta-trained on $\bar{S}$ by performing l optimization steps using the regularized loss

$$L_{\text{reg}}(\theta, \psi) = L(\theta) + R_\psi(\theta)$$

and then evaluated on $\check{S}$ using $L(\theta)$. Since the model parameters adapted during meta-training depend on the regularizer, the meta-test loss also depends on ψ. The parameters ψ can therefore be updated so that the regularizer promotes training updates that generalize across domains.

Over many iterations, the regularizer is learned to penalize classifier parameters in a way that improves cross-domain behaviour. Once R_ψ has been learned, its parameters are frozen, and the final model is trained from scratch on all available source domains using the regularized loss. In this sense, the meta-learning stage does not directly produce the final classifier; it produces a regularization function that is later used during final training.

A direct implementation of this procedure can be computationally expensive, especially when the model contains a deep feature extractor. Training the whole model from scratch inside every meta-learning iteration would be unfeasible for architectures involving high-dimensional image encoders. To reduce this cost, MetaReg decomposes the model into a feature network F_ω and a lighter task network T_θ as follows

$$M_{\omega,\theta} = T_\theta \circ F_\omega$$

This decomposition makes it possible to restrict the inner meta-learning updates to the task-network parameters, while keeping the feature network fixed during the meta-train/meta-test step, considerably reducing the computational overhead.

However, the feature network must still be learned, since the effectiveness of the meta-learning procedure depends on the availability of meaningful latent features across all availability domains.

For this reason, each iteration combines ordinary supervised base-model updates with a meta-learning update of the regularizer.

More precisely, during the regularizer-learning stage, the architecture contains one shared feature network F_ω and one task network T_{θ_i} for each source domain D_i. Each task network is trained only on samples from its corresponding domain, while the feature network is shared across all source domains. At each iteration, the feature-network parameters ω and the domain-specific task-network parameters θ_i are first updated using the ordinary supervised loss, without the learned regularizer. These updates maintain a current set of base parameters $\omega^{(t)}, \theta_1^{(t)}, \ldots, \theta_p^{(t)}$ which are progressively refined throughout training and later provide the starting point for the meta-learning step.

The meta-learning step is then performed by initializing a temporary task network from the current domain-specific task parameters $\theta_i^{(t)}$, rather than from a newly initialized task network. During this step, the feature network $F_{\omega^{(t)}}$ is kept fixed, and only the task-network parameters are adapted on the meta-train domain using the regularized loss. The adapted task network is subsequently evaluated on the meta-test domain, and the resulting (unregularized) meta-test loss is used to update the regularizer parameters. In this way, the feature extractor and the domain-specific task networks provide a continuously improved starting point for meta-learning, while the expensive feature network is not repeatedly retrained inside the inner loop. This considerably reduces the computational overhead.

The same decomposition is natural in the Product-of-Experts model developed in this thesis. The modality-specific encoders and the Product-of-Experts fusion mechanism form the feature component, since they map the available modalities to a latent distribution $x_S \mapsto \hat{p}_\phi(\mathbf{z}|x_S)$. The shared classifier $\hat{p}_\theta(y|\mathbf{z})$ then acts as the task network, since it maps latent representations to diagnostic probabilities. The regularizer $R_\psi(\theta)$ is therefore applied only to the classifier parameters. This is both computationally convenient and conceptually appropriate, because the classifier is precisely the component that must operate across the different latent-input domains.

While this meta-learning framework can be adapted with any parametrized function $R_\psi(\theta)$, in the present work the regularizer is chosen as a constrained variant of the learned weighted L^1 penalty

$$R_\psi(\theta) = \sum_k \text{softplus}(\psi_k)|\theta_k| \tag{4.1}$$

where θ_k denotes the k-th scalar parameter of the classifier. The softplus transformation ensures that the learned regularization weights are non-negative, so that the regularizer remains a proper penalty.

4.2.3. Adaptation to the Product-of-Experts model

We now adapt the meta-regularization framework to the setting of the proposed Product-of-Experts model. The modality-specific regimes D_1, D_2 and $D_{1,2}$ are treated as the available domains for the purpose of meta-learning. These regimes induce three latent-input domains for the shared classifier: an MRI-only domain, where predictions are based on $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$, a PET-only domain, where predictions are based on $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$, and a multimodal domain, where predictions are based on $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$.

During the regularizer-learning stage, we maintain three domain-specific copies of the classifier parameters, denoted by θ_1, θ_2 and $\theta_{1,2}$. These copies are used only for learning the regularizer. After R_ψ has been learned, they are discarded, and the final model is trained from scratch with a single shared classifier $\hat{p}_\theta(y|\mathbf{z})$.

Let $L^{(1)}(\phi, \theta)$, $L^{(2)}(\phi, \theta)$ and $L^{(1,2)}(\phi, \theta)$ denote the supervised losses on D_1, D_2 and $D_{1,2}$, respectively. These losses are computed using the corresponding predictive distributions:

$$L^{(1)}(\phi, \theta) = \sum_{(x_1, y) \in D_1} \mathcal{L}_{\text{CE}}\left(\hat{p}_{\theta,\phi_1}(\cdot|x_1), y\right)$$

$$L^{(2)}(\phi,\theta) = \sum_{(x_2,y)\in D_2} \mathcal{L}_{\mathrm{CE}}\left(\hat{p}_{\theta,\phi_2}(\cdot|x_2), y\right)$$

$$L^{(1,2)}(\phi,\theta) = \sum_{(x_1,x_2,y)\in D_{1,2}} \mathcal{L}_{\mathrm{CE}}\left(\hat{p}_{\theta,\phi}(\cdot|x_1,x_2), y\right)$$

The regularized loss on a regime $r \in \{1, 2, (1,2)\}$ is therefore

$$L^{(r)}_{\mathrm{reg}}(\phi,\theta,\psi) = L^{(r)}(\phi,\theta) + R_\psi(\theta).$$

Because only three regimes are available, the meta-train and meta-test roles can be cycled through rather than sampled randomly. For instance, when D_1 is used as the meta-train regime, the classifier parameters are adapted by performing l gradient steps on $L^{(1)}_{\mathrm{reg}}(\phi,\theta,\psi)$. The adapted classifier is then evaluated on another regime, for example $D_{1,2}$, through $L^{(1,2)}(\phi,\theta^*)$. At the next iterations, different ordered pairs of regimes are used. In this way, the regularizer is optimized to support transfer across the MRI-only, PET-only and multimodal latent-input regimes.

The regularizer learning algorithm proceeds as follows. Each iteration first updates the feature parameters $\phi^{(t)}$ and the domain-specific classifier parameters $\theta_1^{(t)}$, $\theta_2^{(t)}$ and $\theta_{1,2}^{(t)}$ using the ordinary supervised losses

$$L^{(1)}(\phi^{(t)},\theta_1^{(t)}), \qquad L^{(2)}(\phi^{(t)},\theta_2^{(t)}), \qquad L^{(1,2)}(\phi^{(t)},\theta_{1,2}^{(t)}),$$

without the learned regularizer. These updates allow the modality-specific encoders and the Product-of-Experts fusion mechanism to progressively learn meaningful latent representations, while maintaining domain-specific classifier parameters that provide starting points for the subsequent meta-learning step.

The inner meta-learning step is then performed with the feature parameters $\phi^{(t)}$ fixed at their current value. A temporary classifier is initialized from one of $\theta_1^{(t)}$, $\theta_2^{(t)}$ or $\theta_{1,2}^{(t)}$, depending on which regime is selected as meta-train, and is adapted using the corresponding regularized loss

$$L^{(r)}_{\mathrm{reg}}(\phi^{(t)},\theta_r^{(t)},\psi^{(t)}).$$

The adapted classifier is evaluated on a different regime using the ordinary supervised loss, and this meta-test loss is used to update the regularizer parameters $\psi^{(t)}$. Thus, the dependence of the meta-test loss on $\psi^{(t)}$ arises through the regularized inner-loop adaptation of the classifier.

Once the regularizer $R_\psi(\theta)$ has been learned, its parameters are frozen. The full Product-of-Experts model is then trained from scratch using a final objective that includes the supervised classification losses, the learned classifier regularizer and the KL-based latent alignment terms introduced in the following section.

4.3. KL regularization

As motivated earlier, we wish to design a second regularizer that encourages the alignment of the unimodal latent distributions $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$ and $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$ with the multimodal distribution $\hat{p}_\phi(\mathbf{z}|x_1,x_2)$, with the aim of enhancing compatibility between the unimodal and multimodal latent representations. To this end, we introduce a regularization function based on the Kullback-Leibler (KL) divergence, which has been successfully applied in similar multimodal latent variable model settings in the literature [30, 37]. The KL divergence is a scalar quantity defined for a pair of probability distributions $p(z)$ and $q(z)$ as

$$D_{\mathrm{KL}}[p(z)||q(z)] := \int p(z) \log \frac{p(z)}{q(z)} dz.$$

It can be interpreted as a measure of the degree to which the distributions p and q differ. Indeed, while this quantity is not symmetric and thus not a proper distance, it is non-negative and it equals zero if and only if $p(z) = q(z)$ for p-almost all z. Intuitively, this divergence penalizes q for assigning low probability to events that are instead assigned higher probability mass by p.

To be well defined, one must additionally require that p is absolutely continuous with respect to q, meaning that $q(z) = 0 \implies p(z) = 0$. Furthermore, for arbitrary p and q the integral is not solvable analytically and needs to be approximated. Nevertheless, in the special case where p and q are Gaussian, both of the above issues are immediately solved. In particular, since both supports span the whole space $\mathbb{R}^L$, the absolute continuity constraint is automatically satisfied. Secondly, the KL divergence has the following closed-form expression for when $p_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $p_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ [33]:

$$D_{\mathrm{KL}}[p_1(z)||p_2(z)] = \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) - k + \log \frac{|\Sigma_2|}{|\Sigma_1|}\right\}. \tag{4.2}$$

These properties suit our probabilistic model, where $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$ and $\hat{p}_{\phi_i}(\mathbf{z}|x_i)$ are chosen to be multivariate Gaussian distributions.

Motivated by this, we introduce a regularization term that penalizes, on average, the KL divergence from the multimodal latent distribution to each unimodal latent distribution. Concretely, since all three distributions can be computed only for paired samples, we add to the training loss the term

$$\mathcal{L}_{\mathrm{KL}} := \frac{1}{|D_{1,2}|} \sum_{(x_1,x_2)\in D_{1,2}} \left(D_{\mathrm{KL}}\left[\hat{p}_\phi(\mathbf{z}|x_1, x_2) \;||\; \hat{p}_{\phi_1}(\mathbf{z}|x_1)\right] + D_{\mathrm{KL}}\left[\hat{p}_\phi(\mathbf{z}|x_1, x_2) \;||\; \hat{p}_{\phi_2}(\mathbf{z}|x_2)\right]\right). \tag{4.3}$$

This term will then be weighted by λ_{KL}, an hyperparameter controlling the contribution of this regularization. This term corresponds to the empirical training objective used in the experiments: both the MRI-induced and the PET-induced latent distributions are encouraged to remain compatible with the multimodal Product-of-Experts distribution. Only paired samples from $D_{1,2}$ may be used, since otherwise it would not be possible to compute $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$ for the same subject.

It is important to remark that the KL divergence is not symmetric, therefore the choice of the order of the arguments is relevant. In the present setting, we take inspiration from [30], where KL divergence is specifically employed to align latent distribution derived from different modality-patterns.

In our specific setting where the covariance matrices are diagonal a further simplification can be applied, as stated in the following proposition.

Proposition 4.3.1 (KL divergence for Gaussians with diagonal covariance)**.** For $\sigma^2 \in \mathbb{R}^L$ let $\mathrm{diag}(\sigma^2)$ denote the diagonal matrix with the elements σ_i^2 on the diagonal. Then, for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \in \mathbb{R}^L$, we have that

$$D_{\mathrm{KL}}[\mathcal{N}(\mu_1, \mathrm{diag}(\sigma_1^2))||\mathcal{N}(\mu_2, \mathrm{diag}(\sigma_2^2))] = \frac{1}{2}\sum_{i=1}^{L}\left(\frac{\sigma_{1,i}^2}{\sigma_{2,i}^2} + \frac{(\mu_{2,i} - \mu_{1,i})^2}{\sigma_{2,i}^2} + \log\frac{\sigma_{2,i}^2}{\sigma_{1,i}^2} - 1\right)$$

Proof. Letting $\Sigma_1 = \mathrm{diag}(\sigma_1^2)$ and $\Sigma_2 = \mathrm{diag}(\sigma_2^2)$, the following identities hold:

$$\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) = \sum_{i=1}^{L}\frac{\sigma_{1,i}^2}{\sigma_{2,i}^2}$$

$$(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) = \sum_{i=1}^{L} \frac{(\mu_{2,i} - \mu_{1,i})^2}{\sigma_{2,i}^2}$$

$$\log \frac{|\Sigma_2|}{|\Sigma_1|} = \log \frac{\prod_{i=1}^{L} \sigma_{2,i}^2}{\prod_{j=1}^{L} \sigma_{1,j}^2} = \sum_{i=1}^{L} \log \frac{\sigma_{2,i}^2}{\sigma_{1,i}^2}$$

Substituing these expressions in (4.2) and rearranging the terms in a single summation yields the desired expression. □

As a consequence of this proposition, denoting

$$\hat{p}_{\phi_1}(\mathbf{z}|x_1) \sim \mathcal{N}(\mu_1, \mathrm{diag}(\sigma_1^2)), \quad \hat{p}_{\phi_2}(\mathbf{z}|x_2) \sim \mathcal{N}(\mu_2, \mathrm{diag}(\sigma_2^2)), \quad \hat{p}_{\phi}(\mathbf{z}|x_1, x_2) \sim \mathcal{N}(\mu_{1,2}, \mathrm{diag}(\sigma_{1,2}^2)).$$

the two KL terms reduce to

$$D_{\mathrm{KL}}\left[\hat{p}_{\phi}(\mathbf{z}|x_1, x_2) \,||\, \hat{p}_{\phi_1}(\mathbf{z}|x_1)\right] = \frac{1}{2} \sum_{i=1}^{L} \left(\frac{\sigma_{1,2,i}^2}{\sigma_{1,i}^2} + \frac{(\mu_{1,i} - \mu_{1,2,i})^2}{\sigma_{1,i}^2} + \log \frac{\sigma_{1,i}^2}{\sigma_{1,2,i}^2} - 1 \right),$$

and

$$D_{\mathrm{KL}}\left[\hat{p}_{\phi}(\mathbf{z}|x_1, x_2) \,||\, \hat{p}_{\phi_2}(\mathbf{z}|x_2)\right] = \frac{1}{2} \sum_{i=1}^{L} \left(\frac{\sigma_{1,2,i}^2}{\sigma_{2,i}^2} + \frac{(\mu_{2,i} - \mu_{1,2,i})^2}{\sigma_{2,i}^2} + \log \frac{\sigma_{2,i}^2}{\sigma_{1,2,i}^2} - 1 \right).$$

These expressions can now be readily applied in the training of the model, which we describe in the next section by illustrating the detailed algorithms.

4.4. Training algorithms

The complete training procedure is divided into two stages. In the first stage, the parameters ψ of the classifier regularizer are learned using the meta-regularization procedure described above. During this stage, three domain-specific copies of the classifier, with parameters θ_1, θ_2 and $\theta_{1,2}$, are maintained only for the purpose of learning the regularizer. These copies are not part of the final model. The feature parameters ϕ are updated through ordinary supervised losses, while the inner meta-learning step adapts only a temporary classifier initialized from one of the domain-specific copies. The meta-test loss is then used to update ψ, so that the learned regularizer promotes classifier updates that transfer across the MRI, FDG-PET and MRI+FDG-PET specific regimes.

Algorithm 1 Meta-learning of the classifier regularizer

Require: Datasets D_1, D_2, $D_{1,2}$
Require: Number of iterations N_{iter}, inner steps l
Require: Learning rates α_1, α_2
Require: Initial parameters $\phi^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \theta_{1,2}^{(0)}, \psi^{(0)}$
for $t = 1, \ldots, N_{\text{iter}}$ **do**
Base-model update
  for $r \in \{1, 2, (1,2)\}$ **do**
    Sample mini-batch $\mathcal{B}_r \subset D_r$
  end for
  $\phi^{(t)} \leftarrow \phi^{(t-1)} - \alpha_1 \nabla_\phi \sum_{r \in \{1,2,(1,2)\}} L_{\mathcal{B}_r}^{(r)} \left(\phi^{(t-1)}, \theta_r^{(t-1)} \right)$
  for $r \in \{1, 2, (1,2)\}$ **do**
    $\theta_r^{(t)} \leftarrow \theta_r^{(t-1)} - \alpha_1 \nabla_{\theta_r} L_{\mathcal{B}_r}^{(r)} \left(\phi^{(t-1)}, \theta_r^{(t-1)} \right)$
  end for
Meta-train/meta-test split
  Select two distinct regimes $a, b \in \{1, 2, (1,2)\}$ with $a \neq b$
Meta-train
  Initialize temporary classifier $\beta_1 \leftarrow \theta_a^{(t)}$
  for $s = 2, \ldots, l+1$ **do**
    Sample meta-train mini-batch $\mathcal{B}_a^{\text{tr}} \subset D_a$
    $\beta_s \leftarrow \beta_{s-1} - \alpha_2 \nabla_{\beta_{s-1}} \left[L_{\mathcal{B}_a^{\text{tr}}}^{(a)} \left(\phi^{(t)}, \beta_{s-1} \right) + R_{\psi^{(t-1)}}(\beta_{s-1}) \right]$
  end for
  $\widehat{\theta}_a^{(t)} \leftarrow \beta_l$
Meta-test
  Sample meta-test mini-batch $\mathcal{B}_b^{\text{te}} \subset D_b$
  Evaluate the ordinary supervised loss $L_{\mathcal{B}_b^{\text{te}}}^{(b)} \left(\phi^{(t)}, \widehat{\theta}_a^{(t)} \right)$
Meta-update
  $\psi^{(t)} \leftarrow \psi^{(t-1)} - \alpha_2 \nabla_\psi L_{\mathcal{B}_b^{\text{te}}}^{(b)} \left(\phi^{(t)}, \widehat{\theta}_a^{(t)} \right) \Big|_{\psi = \psi^{(t-1)}}$
end for
return learned regularizer parameters $\psi^{(N_{\text{iter}})}$

After the regularizer has been learned, its parameters are frozen and the final Product-of-Experts model is trained from scratch. In contrast to the first stage, this model contains a single shared classifier $\hat{p}_\theta(y|\mathbf{z})$, as defined in Chapter 3. Training involves samples from the MRI-specific training set D_1, the FDG-PET-specific training set D_2, and the complete multimodal training set $D_{1,2}$. In addition, complete samples are used to compute the KL alignment term, since only for these samples are $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$, $\hat{p}_{\phi_1}(\mathbf{z}|x_1)$ and $\hat{p}_{\phi_2}(\mathbf{z}|x_2)$ available for the same subject.

Algorithm 2 Final training of the Product-of-Experts model

Require: Datasets D_1, D_2, $D_{1,2}$
Require: Learned regularizer parameters ψ^*
Require: Learning rate η, number of iterations N_{train}
Require: Hyperparameters $\lambda_1, \lambda_2, \lambda_{1,2}, \lambda_R, \lambda_{\text{KL}}$
Require: Initial model parameters $\phi^{(0)}, \theta^{(0)}$

for $t = 1, \ldots, N_{\text{train}}$ **do**
  Sample mini-batch $\mathcal{B}_1 \subset D_1$
  Sample mini-batch $\mathcal{B}_2 \subset D_2$
  Sample mini-batch $\mathcal{B}_{1,2} \subset D_{1,2}$

Supervised classification losses

  Compute the MRI-only loss

$$L_{\mathcal{B}_1}^{(1)} = \frac{1}{|\mathcal{B}_1|} \sum_{(x_1,y)\in\mathcal{B}_1} \mathcal{L}_{\text{CE}} \left(\hat{p}_{\theta^{(t-1)},\phi_1^{(t-1)}}(\cdot|x_1), y \right)$$

  Compute the PET-only loss

$$L_{\mathcal{B}_2}^{(2)} = \frac{1}{|\mathcal{B}_2|} \sum_{(x_2,y)\in\mathcal{B}_2} \mathcal{L}_{\text{CE}} \left(\hat{p}_{\theta^{(t-1)},\phi_2^{(t-1)}}(\cdot|x_2), y \right)$$

  Compute the multimodal loss

$$L_{\mathcal{B}_{1,2}}^{(1,2)} = \frac{1}{|\mathcal{B}_{1,2}|} \sum_{(x_1,x_2,y)\in\mathcal{B}_{1,2}} \mathcal{L}_{\text{CE}} \left(\hat{p}_{\theta^{(t-1)},\phi^{(t-1)}}(\cdot|x_1,x_2), y \right)$$

KL term

  Compute the multimodal-to-unimodal alignment loss on paired samples

$$L_{\text{KL},\mathcal{B}_{1,2}} = \frac{1}{|\mathcal{B}_{1,2}|} \sum_{(x_1,x_2,y)\in\mathcal{B}_{1,2}} \left(D_{\text{KL}} \left[\hat{p}_{\phi^{(t-1)}}(\mathbf{z}|x_1,x_2) \| \hat{p}_{\phi_1^{(t-1)}}(\mathbf{z}|x_1) \right] + D_{\text{KL}} \left[\hat{p}_{\phi^{(t-1)}}(\mathbf{z}|x_1,x_2) \| \hat{p}_{\phi_2^{(t-1)}}(\mathbf{z}|x_2) \right] \right)$$

Total objective

  Form the mini-batch objective

$$\mathcal{J}_{\mathcal{B}_1,\mathcal{B}_2,\mathcal{B}_{1,2}} = \lambda_1 L_{\mathcal{B}_1}^{(1)} + \lambda_2 L_{\mathcal{B}_2}^{(2)} + \lambda_{1,2} L_{\mathcal{B}_{1,2}}^{(1,2)} + \lambda_R R_{\psi^*}(\theta^{(t-1)}) + \lambda_{\text{KL}} L_{\text{KL},\mathcal{B}_{1,2}}.$$

Parameter update

  $(\phi^{(t)}, \theta^{(t)}) \leftarrow (\phi^{(t-1)}, \theta^{(t-1)}) - \eta \nabla_{\phi,\theta} \mathcal{J}_{\mathcal{B}_1,\mathcal{B}_2,\mathcal{B}_{1,2}}$

end for

return trained parameters $\phi^{(N_{\text{train}})}, \theta^{(N_{\text{train}})}$

5

Experimental Setup

This chapter describes the experimental setup used throughout this thesis. Section 5.1 presents the original ADNI data and the preprocessing pipeline applied to it, concluding with the final composition of the dataset and the training, validation and test splits. Section 5.2 describes the architecture of the Product-of-Experts model, including the modality-specific encoders and the shared classifier. Section 5.3 introduces the deterministic and stochastic approximations used to compute the predictive distributions. Section 5.4 defines the evaluation metrics used to assess model performance. Finally, Section 5.5 describes the design of the experiments, including the model selection procedure and the specific comparisons carried out in this thesis.

5.1. Dataset and Preprocessing

This section describes the preprocessing pipeline and the composition of the dataset used in the experiments. A short overview of the original ADNI data is given first, motivating the preprocessing steps described afterward. The final composition of the dataset resulting from this procedure is then presented. Finally, the cross-validation scheme adopted for the experiments is described, together with the method used to perform subject-level splitting.

5.1.1. ADNI original data

The full ADNI cohort contains clinical information and imaging data from more than 3300 subjects, collected over 21 years across five main phases. With the primary goal of investigating biomarkers of the disease, the study includes several data types from different modalities, such as cerebrospinal fluid measurements, genetic information, cognitive scores, T1 and T2-weighted MRI scans, as well as PET acquisitions with multiple radiotracers. In this thesis, we consider exclusively diagnostic assessments, T1-weighted MRI acquisitions and FDG-PET acquisitions.

Diagnostic assessments divide patients into three categories: cognitively normal (CN), mild cognitive impairment (MCI), or Alzheimer's disease (AD). These diagnoses are based exclusively on clinical and cognitive assessments, including tools such as the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating (CDR), and the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog).

The study follows a longitudinal design, meaning that patients are followed over time rather than assessed only once. The length of this follow-up period is generally at least one year, although it can vary depending on the initial diagnostic category of the patient and on the specific phase of the study. During this period, the diagnostic assessments, MRI and FDG-PET acquisitions are performed on separate visits, which are often several months apart or more. This temporal misalignment represents a central difficulty, both in assigning diagnostic labels to the acquisitions and in constructing multimodal samples, since the diagnosis and the neuroimaging acquisitions

should ideally describe the same clinical state of the subject.

Furthermore, the raw MRI and FDG-PET volumes are not directly suited for training deep learning models, since factors such as differences in head position or the presence of non-brain tissues introduce statistical variability that is unrelated to disease progression. This additional variability risks confounding the model, as it may obscure the patterns that are actually relevant to AD.

For these reasons, several preprocessing steps are required in order to obtain the final dataset. Specifically, all the MRI scans are first transformed using a Matlab pipeline, which removes all the non-brain tissues and aligns the anatomical structures to a common template space. A similar procedure is also applied to the FDG-PET volumes. Each MRI and FDG-PET scan is then assigned a diagnosis label (CN, MCI or AD) based on a temporal proximity criterion. Finally, multimodal samples are created by pairing temporally close MRI and FDG-PET scans of the same patient with the same diagnosis. These steps are illustrated in more detail in the following sections.

5.1.2. MRI preprocessing

The MRI scans downloaded from ADNI are stored as DICOM series, where each scan consists of a sequence of two dimensional slices. The first step is to convert each DICOM series into a three dimensional NIfTI volume.

The main preprocessing operations are then performed using MATLAB scripts based on the Statistical Parametric Mapping (SPM) library [38]. For each raw MRI volume, SPM applies bias correction, tissue segmentation, and spatial normalization within a unified probabilistic framework. Bias correction reduces low frequency intensity artifacts caused by inhomogeneities in the magnetic field of the scanner. Tissue segmentation assigns to each voxel a probability of belonging to different tissue classes, including gray matter, white matter, cerebrospinal fluid, skull, and other non brain tissues. Spatial normalization estimates a smooth transformation that maps the subject specific volume to the standard MNI152 template space [39]. This transformation is represented by a deformation field, denoted by y.

The output of this step consists of the MRI volume in MNI152 space, the tissue probability maps, and the deformation field y. The tissue probability maps are then used to create a binary brain mask. In particular, voxels whose probability of belonging to gray matter or white matter exceeds a fixed threshold are retained, while the remaining voxels are removed. Applying this mask suppresses the skull and other non brain tissues, and is therefore referred to as skull stripping.

The resulting MRI volume is a brain extracted NIfTI image of size $91 \times 109 \times 91$ in MNI152 space. Voxels outside the brain mask are set to zero, so that non brain regions do not contribute to the model. Finally, z score normalization is applied to the intensities inside the brain mask, in order to make the intensity distributions more comparable across subjects. An example preprocessed volume is shown in Figure 5.1.

5.1.3. FDG-PET preprocessing

The FDG-PET scans require a related but slightly different preprocessing procedure. As with MRI, the raw PET volumes are affected by differences in head position and acquisition geometry, and must be brought into a common anatomical space. However, PET images contain substantially less anatomical detail than MRI images, since their signal reflects the distribution and activity of the injected radiotracer rather than structural tissue contrast. As a consequence, directly normalizing PET volumes to a standard anatomical template is less reliable. For this reason, the PET preprocessing makes use of the deformation fields obtained from the MRI preprocessing.

Each PET acquisition is first converted from DICOM to NIfTI format. In contrast to MRI, the converted PET file may be four dimensional, because the radiotracer signal is sampled over multiple time frames during the acquisition. These frames are not necessarily aligned with each other, so they are first coregistered. A temporal mean is then computed, yielding a single three

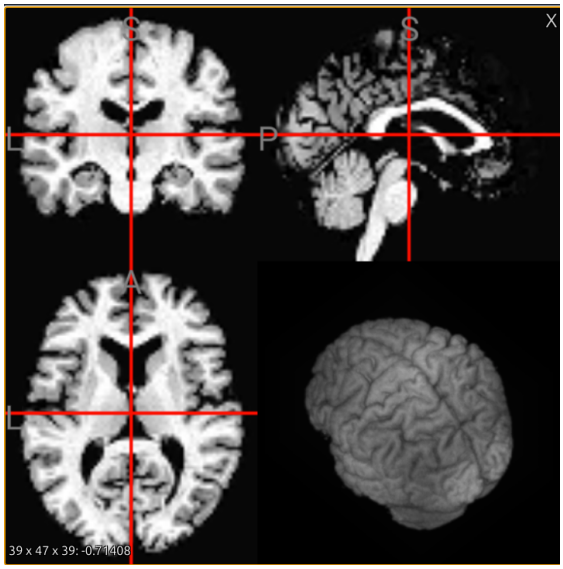


Figure 5.1: Preprocessed MRI volume

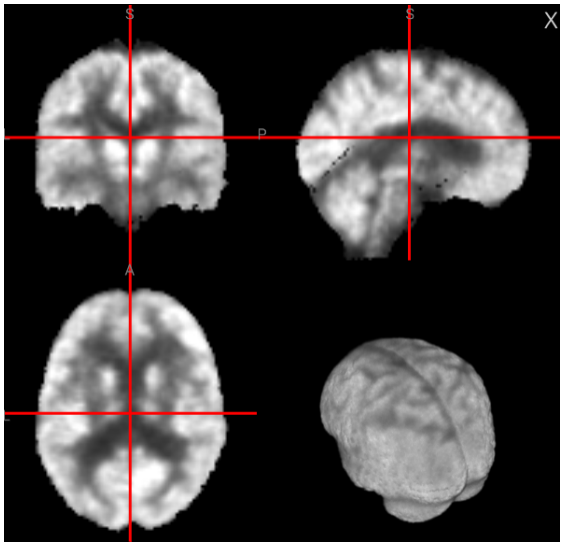


Figure 5.2: Preprocessed FDG-PET volume

dimensional PET volume for the acquisition.

To map the PET volume to MNI152 space, each PET scan is paired with an MRI scan from the same subject. At this stage, the purpose of the pairing is only to exploit the anatomical structure of the MRI scan, so the MRI and PET acquisitions do not need to satisfy the temporal matching criterion used later for sample construction. The PET volume is first rigidly registered to the corresponding MRI volume. Once the two scans are aligned in the subject specific space, the MRI deformation field y is applied to the PET volume, thereby warping it to the MNI152 template.

The same brain mask obtained during MRI preprocessing is then applied to the PET volume, so that signal outside the brain is removed. Finally, z score normalization is applied to the resulting PET volume. The final preprocessed FDG-PET scan has the same spatial dimensions as the MRI volumes, namely $91 \times 109 \times 91$. An example preprocessed FDG-PET volume is shown in Figure 5.2.

5.1.4. Label matching

After the MRI and FDG-PET scans have been preprocessed, the next step is to assign a diagnostic label to each individual scan. Ideally, each MRI and FDG-PET scan should be assigned a diagnosis that refers to exactly the same clinical state of the subject. In practice, as previously explained, this is not possible in ADNI, because the neuroimaging modalities and the diagnostic evaluations are often acquired at different time points. For this reason, assigning a diagnosis to each scan requires a temporal matching procedure. We assume that, within a sufficiently contained time window Δ, the disease state of the subject does not change substantially, and we therefore match scans and diagnostic visits when their acquisition dates are close enough. To be able to obtain a sufficiently large number of samples, we set $\Delta = 180$ days. Although such large time window is necessary due to modalities and diagnoses being acquired often several months apart, it also represent an important limitation, since these large time differences may violate our diagnosis consistency assumption.

Each acquisition is associated with a unique identifier, through which its acquisition date can be recovered. For each scan, the diagnostic records of the corresponding subject are searched, and all diagnostic visits within the temporal window Δ from the acquisition date are retained. If one or more such visits exist, the closest diagnostic visit in time is selected, and its diagnosis is assigned to the scan.

There are cases in which no diagnostic visit falls within the window Δ. In these cases, a secondary rule is applied. If the closest diagnostic visit before the scan and the closest diagnostic visit after the scan exist and have the same diagnosis, then this diagnosis is assigned to the scan, even if the two visits are outside the window Δ. The motivation is that if the subject has the same diagnosis immediately before and immediately after the scan, it is reasonable to assume that the diagnosis at the acquisition time was unchanged. If neither the temporal window rule nor this secondary rule can assign a label, the scan is excluded from the dataset. At the end of this step, each retained MRI or FDG-PET scan has an associated diagnostic label.

5.1.5. Multimodal samples creation

The remaining task is to combine individual scans into multimodal observations. This is particularly challenging because MRI and FDG-PET acquisitions for the same subject are generally not taken at the same time, even though they should ideally reflect the same clinical state of the patient. As in the label matching step, we assume that the disease state of a patient does not change substantially within the same temporal window $\Delta = 180$ days. Under this assumption, the task of creating multimodal samples reduces to matching MRI and FDG-PET acquisitions that belong to the same subject and share the same diagnosis, while respecting the temporal window and ensuring that each individual scan is used in at most one sample.

To see why a naive matching rule can be suboptimal, consider the example illustrated in Figure 5.3 with two MRI scans, M_1 and M_2, and two PET scans, P_1 and P_2, all from the same subject and with the same assigned diagnosis. Suppose that M_2 lies within the time window of both P_1 and P_2, whereas M_1 lies within the time window of P_1 only. If a greedy algorithm assigns M_2 to P_1, for example because they are closest in time, then P_1 and M_2 remain unmatched, and only one complete multimodal sample is obtained. However, the better solution is to match M_2 with P_2 and M_1 with P_1, producing two complete samples. This example shows that the matching procedure must consider the global structure of possible pairings, rather than making only local decisions.

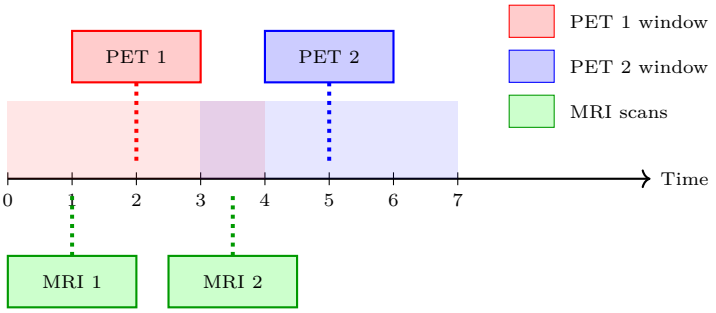


Figure 5.3: Illustration of the mismatched scans example. If PET 1 is matched with MRI 2, the other scans remain unmatched.

For this reason, a heuristic matching algorithm is used. The algorithm was initially designed for a more general setting with $M \geq 2$ modalities, although in the present work it is applied only to MRI and FDG-PET. For a fixed subject and diagnosis, let $m_1, \ldots, m_M$ denote the available modality groups. The algorithm first selects the modality with the smallest number of observations, denoted by $\hat{m}$, and uses it as the anchor modality. This choice is natural because the number of disjoint complete samples cannot exceed the number of observations available in the least represented modality.

For each observation j of the anchor modality $\hat{m}$, the algorithm identifies the observations from the remaining modalities whose acquisition dates differ from that of j by at most Δ. The possible multimodal samples associated with j are obtained by taking all combinations of the admissible observations from the other modalities. If $S_i^{(j)}$ denotes the set of observations of modality i lying within the time window of anchor observation j, then the candidate set for j is given by

$$S^{(j)} = S_{i_1}^{(j)} \times \ldots \times S_{i_{M-1}}^{(j)},$$

where $i_1, \ldots, i_{M-1}$ denote the non anchor modalities. The construction also allows one modality to be missing, so that incomplete samples can be generated when a complete match is not available.

The full set of candidate samples is then $S = \cup_{j \in \hat{m}} S^{(j)}$. The goal is to select a subset of S such that no original scan appears in more than one selected sample. To achieve this, each candidate sample $s_k \in S$ is assigned a weight w_k equal to the number of modalities present in that sample. Boolean variables $y_k \in \{0, 1\}$ indicate whether candidate sample s_k is selected. The selection is then formulated as an Integer Linear Programming problem with objective

$$\sum_{k \in S} y_k w_k,$$

subject to the constraint that each individual scan can be selected at most once. In particular, this constraint is encoded as follows: for each pair of non-disjoint samples k_1 and k_2, we introduce the clause $y_{k_1} \implies \neg y_{k_2}$, where $\implies$ denotes logical implication and $\neg$ denotes logical negation. This ensures that if k_1 is selected, k_2 is not. The constraint works symmetrically: if $y_{k_2} = 1$, then $\neg y_{k_2} = 0$, which in turn forces $y_{k_1} = 0$. Therefore, a single clause suffices for each unordered pair $\{k_1, k_2\}$, resulting in at most $\frac{1}{2}|S|(|S| - 1)$ constraints.

This objective favours samples containing more modalities, while still allowing incomplete samples when no complete match is possible. Because incomplete candidates are included, the algorithm always returns one sample for each observation of the anchor modality. Once these samples have been selected, the corresponding original observations are removed from the available pool, and the algorithm can be repeated without the previous anchor modality if needed.

In the present work, the algorithm is applied with $M = 2$, corresponding to MRI and FDG-PET. Since MRI scans are considerably more abundant than FDG-PET scans, the procedure matches

Table 5.1: Dataset composition by pairing status and diagnosis.

	CN	MCI	AD	Total
D^m (Unpaired MRI)	3014	3885	1739	8638
D^p (Unpaired FDG-PET)	122	148	104	374
D^f (Complete)	538	1119	453	2110
Total	3674	5152	2296	11122

most PET scans with an MRI scan and produces the complete dataset D^f of samples (x_1, x_2, y). The remaining unpaired MRI and FDG-PET scans form instead the unpaired samples subsets D^m and D^p.

5.1.6. Final dataset

The final dataset resulting from the preprocessing pipeline is composed of the unpaired MRI samples (D^m), unpaired FDG-PET samples (D^p) and complete samples (D^f), as defined in Definition 3.1.1. The histogram in Figure 5.4 highlights the pairing-status imbalance at the core of this thesis: 8638 unpaired MRI samples constitute the vast majority of the dataset, whereas complete samples account for roughly one fifth of the dataset, with 2110 samples. Unpaired FDG-PET samples are present in very small numbers, accounting for only 374 samples.

The horizontal bar plot in Figure 5.4 depicts the relative proportions of diagnostic classes within each of the three sets. MCI remains the most represented class in all cases, followed by CN and AD, although the diagnostic composition varies to some extent across the three sets. In the complete samples, MCI accounts for up to 53% of samples, with the remainder split roughly evenly between CN (25%) and AD (21%). The opposite extreme is found in the unpaired FDG-PET samples, where MCI accounts for only 40%, while CN and AD rise to 33% and 28%, respectively.

Since the number of samples is limited, each modality-availability pathway (MRI-available, FDG-PET-available, and both MRI and FDG-PET available) is trained using every available acquisition of the relevant modality. Recall from Definition 3.1.2 that the MRI-available set D_1 combines all MRI acquisitions from D^m and D^f, while the FDG-PET-available set D_2 combines all FDG-PET acquisitions from D^p and D^f. The multimodal set $D_{1,2}$ coincides with the complete samples set D^f. In other words, the MRI and FDG-PET acquisitions belonging to complete samples are reused two times: once as part of D^f, and again as part of D_1 or D_2. The number of underlying unique acquisitions remains therefore unchanged. The modality availability sets D_1 and D_2, along with the complete set $D_{1,2} := D^f$ are ultimately used to train the models. The final dataset employed during the experiments can therefore be expressed as

$$D := D_1 \cup D_2 \cup D_{1,2}$$

The composition of these sets is shown in Figure 5.5 and Table 5.2. Because the number of unpaired FDG-PET samples is small, combining them with the acquisitions from the complete samples results in $D_{1,2}$ and D_2 having a very similar diagnostic composition. The MRI-available set D_1, in contrast, still shows some variation relative to D^f: the proportion of CN samples increases slightly and the proportion of MCI samples decreases slightly, while the proportion of AD samples remains roughly constant across all three sets. This variation in diagnostic composition across modality-availability sets should also be reflected during the training, validation and test phases, motivating the stratified group splitting procedure described in the next subsection.

5.1.7. Training, validation and test splits

To train, validate and evaluate the model, a 5-fold cross-validation procedure with a held-out test set is employed. Each experiment is repeated 5 times: at each repetition, 4 of the 5 folds are used as the training set, while the remaining fold serves as the validation set. For each fold, the epoch achieving the highest validation balanced accuracy is selected, giving one trained model per fold.

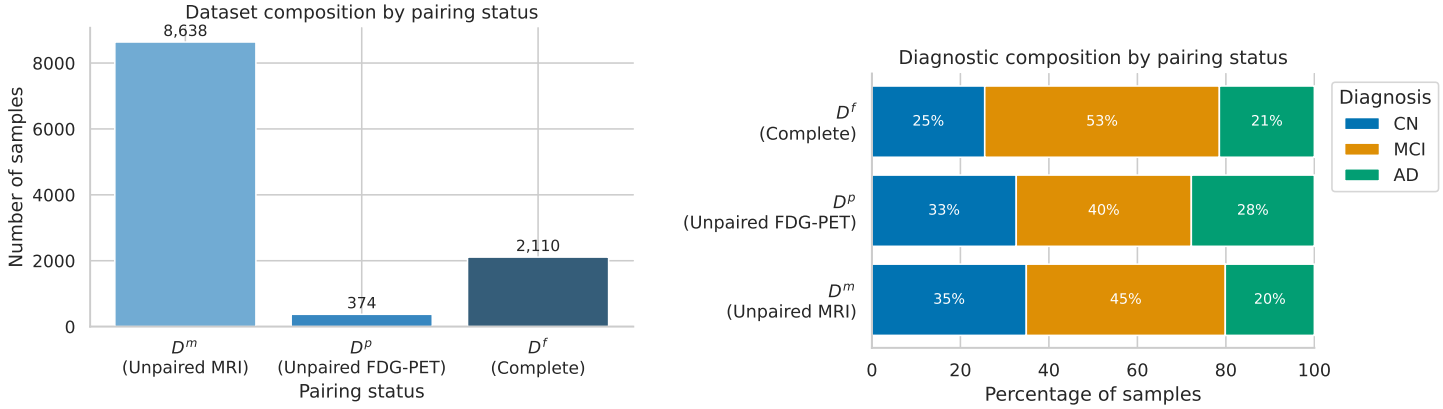


Figure 5.4: Unpaired MRI samples (D^m), unpaired FDG-PET samples (D^p), and complete samples (D^f): set sizes (left) and relative diagnostic class composition (right).

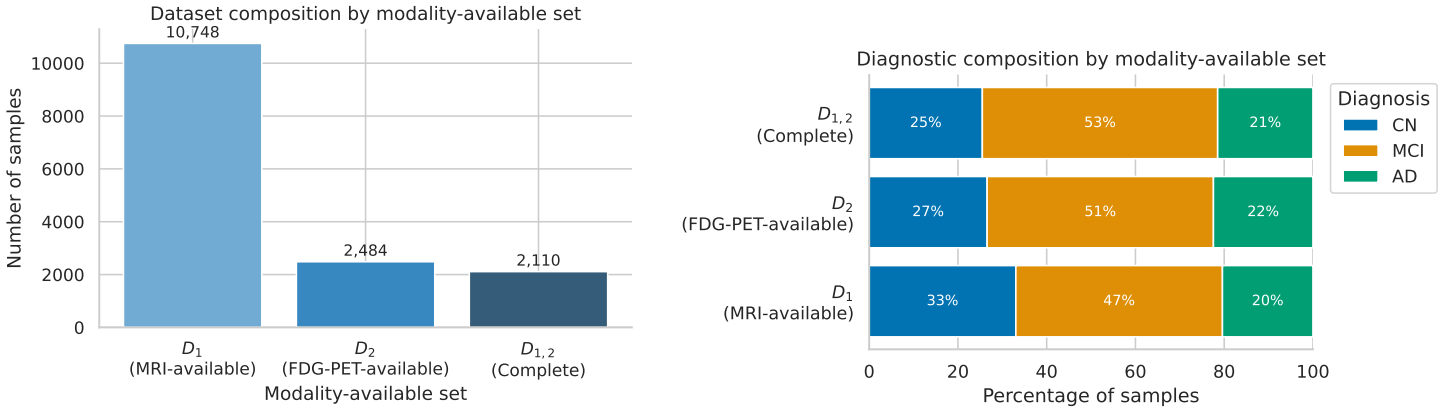


Figure 5.5: Number of samples available to the MRI-available, FDG-PET-available and multimodal pathways (left) and their relative diagnostic class composition (right). Since complete samples contribute to more than one pathway, the sets shown are not mutually exclusive.

Table 5.2: Dataset composition by modality availability and diagnosis.

	CN	MCI	AD	Total
D_1 (MRI-available)	3552	5004	2192	10748
D_2 (FDG-PET-available)	660	1267	557	2484
$D_{1,2}$ (Complete)	538	1119	453	2110

The held-out test set is used exclusively for final evaluation, and is never involved in any form of parameter or configuration choice. In particular, whenever an experiment requires comparing several configurations, such as different training objectives or regularization strengths, this comparison is always based on validation performance alone. Only after a configuration has been selected in this way are the corresponding five fold-specific models evaluated on the held-out test set, and the resulting five values are reported as mean and standard deviation. This ensures that the test-set results reported throughout this chapter reflect the generalization performance of already-fixed configurations, rather than being influenced by the process used to select them.

This approach requires the dataset to be divided into 6 distinct partitions: 5 cross-validation folds and a held-out test set. The held-out test set is selected first, corresponding to approximately 15% of the complete dataset. The remaining 85% is then split into the 5 cross validation folds. Both the test set and the 5 folds are created once, before any experiment is run, and remain fixed throughout the entire study. These partitions are depicted in Figure 5.6. Recall that the dataset used in the experiment is $D := D_1 \cup D_2 \cup D_{1,2}$, as specified in the earlier section.

When creating the partitions, two aspects need to be taken into account. First, the training, validation, and test sets should contain similar proportions of diagnostic classes, as well as of complete and unpaired samples. Second, and more importantly, all samples belonging to a given subject must be assigned to a single partition, whether training, validation, or test. This is necessary to avoid a phenomenon known as data leakage: since MRI and FDG-PET scans encode structural information about the brain that is specific to each subject, a model could learn to recognize individual patients rather than disease-relevant patterns. If scans from the same subject appeared in both the training and test sets, the model might correctly predict the diagnosis simply because it had already learned to associate that particular brain structure with that patient during training, leading to unrealistically optimistic results.

To address both requirements at once, the dataset is partitioned using a stratified group splitting approach. For each sample, a stratification key is defined, encoding both the diagnosis and its pairing status: whether it is unpaired MRI, unpaired FDG-PET or a complete sample. In parallel, each subject defines a group, containing all samples associated with that subject. The splitting algorithm then ensures that all samples within the same group are assigned to the same partition, while trying to keep the distribution of stratification keys as balanced as possible across partitions. In this way, data leakage is avoided, while diagnostic classes and pairing status remain well balanced across training, validation, and test sets.

5.1.8. Augmentation

A major issue arising when highly parametrized models are trained on relatively small datasets is overfitting. A classic approach to this problem is to apply augmentation transforms to the training data. The end goal is to add artificial variability in the data to simulate new samples from the true data distribution by means of small transformations that change the voxels and effectively create a new sample, without altering underlying disease related signals. In the present setting, we apply small translations, rotations and scaling along different axis, as well as Gaussian noise and reflection along the horizontal axis, which preserves real anatomical patterns. Each transformation is unique to each sample, since its parameters (e.g. the magnitude of the translation, rotation and scaling) are sampled from specific distributions. Moreover, not all transforms are applied at the same time; rather, each occurs independently with a given probability p_t specific to each transform t. An optimal configuration of the augmentation transforms is found empirically and is described in Table 5.3. In multimodal samples, the same transformation is applied to both modalities. An example of the effect of the augmentation transform is depicted in Figure 5.7.

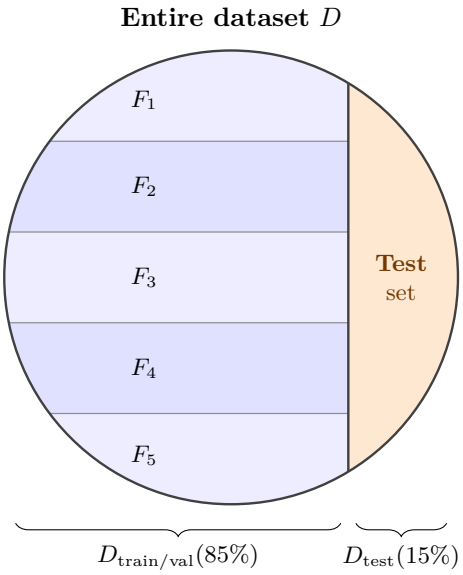


Figure 5.6: Partitions of the dataset. A held out test set (right) is first selected, while the remaining training and validation set (left) is further divided into 5 folds.

Table 5.3: Data augmentation transforms applied during training. Each transform is applied independently with probability p. For multimodal samples, the same spatial transform is applied to both modalities.

Transform	Probability p	Parameterization
Random flip	0.5	Applied along a randomly selected spatial axis
Random affine	0.5	Rotation angle sampled from $\mathcal{U}(-5°, 5°)$ Translation offset sampled from $\mathcal{U}(-5, 5)$ voxels along each axis
Random Gaussian noise	0.2	Standard deviation sampled from $\mathcal{U}(0, 0.1)$
Gaussian blur	1.0	Standard deviation fixed to $\sigma = 0.8$

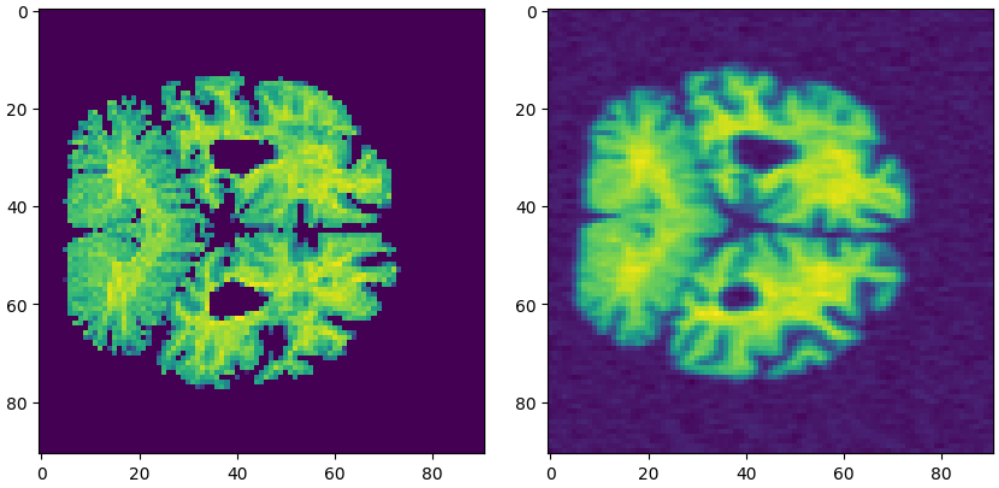


Figure 5.7: Coronal slice of an MRI volume before (left) and after (right) augmentation. The transform applies a fixed Gaussian blur to all volumes, while in this particular example also a slight clockwise rotation along the sagittal axis and some Gaussian noise are visible.

5.2. Model architecture

Recall the Product-of-Experts model defined in Section 3. For any available modality set $S \subseteq \{1, 2\}$, the model first constructs a latent conditional distribution

$$\hat{p}_\phi(\mathbf{z}|x_S) \propto p(\mathbf{z}) \prod_{i \in S} \tilde{p}_{\phi_i}(\mathbf{z}|x_i),$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$ is the prior and each auxiliary expert is $\tilde{p}_{\phi_i}(\mathbf{z}|x_i) = \mathcal{N}(\mathbf{z}; \mu_i, \Sigma_i)$. The diagnostic predictive distribution is then defined as

$$\hat{p}_{\theta,\phi}(y|x_S) = \mathbb{E}_{\hat{p}_\phi(\mathbf{z}|x_S)} [\hat{p}_\theta(y|\mathbf{z})]$$

Therefore, the learnable components of the model are the following. First, two modality-specific encoders map the corresponding modality x_i (when available) to the parameters μ_i and Σ_i of the auxiliary Gaussian expert $\tilde{p}_{\phi_i}(\mathbf{z}|x_i)$. Second, a shared classifier $\hat{p}_\theta(y|\mathbf{z})$ maps instead the latent representation $\mathbf{z}$ to a discrete probability distribution over the possible diagnoses.

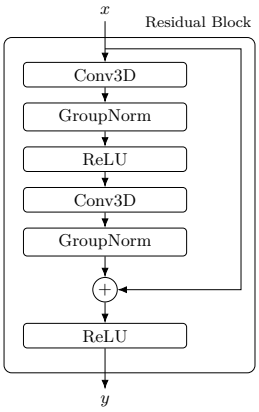
5.2.1. Modality-specific encoders

Formally, a modality specific encoder is a function

$$\begin{aligned} E^i_{\phi_i} : \mathcal{X} &\longrightarrow \mathbb{R}^L \times \mathbb{R}^L, \\ x_i &\longmapsto (\mu_i, \Sigma_i) \end{aligned}$$

which maps the modality x_i to the parameters $\mu_i \in \mathbb{R}^L$ and $\Sigma_i \in \mathbb{R}^L$ of the auxiliary Gaussian expert $\tilde{p}_{\phi_i}(\mathbf{z}|x_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$. For simplicity, the covariance matrices Σ_i are chosen to be diagonal, therefore only L variances $\sigma^2_{i,j}$ with $j \in \{1, ..., L\}$ are needed to encode it, L denoting the dimensionality of the latent variable $\mathbf{z} \in \mathbb{R}^L$. For our experiment, the latent dimensionality was set to $L = 256$.

The encoders are implemented as 3-dimensional convolutional neural networks (3D CNNs). Each convolutional layer convolves the input features with kernels to extract higher level features. Max pooling layers are instead introduced to gradually reduce the dimensionality of the features. Further, activation layers apply the non-linear activation function ReLU, whereas group normalization layers normalize the latent features in order to facilitate learning [40]. In the encoder architecture, these layers are joined into higher level groups called Residual Blocks, illustrated in Figure 5.8, together with the parameters for each of its layer. Every residual block has two parameters: the output channels C_{out}, and the stride s. Normally $s = 1$, and the spatial dimension of the input remain unchanged; however, it is possible to set $s = 2$, which halves the spatial dimensions of the features: in this sense, the residual block is able to act similarly to a pooling layer. The skip connection is implemented as a $1 \times 1 \times 1$ convolution, and it helps preventing the gradients from deep layers from vanishing [41].



Layer	Kernel	Stride	Output channels
Conv3D	$3 \times 3 \times 3$	s	C_{out}
GroupNorm	–	–	C_{out}
ReLU	–	–	C_{out}
Conv3D	$3 \times 3 \times 3$	1	C_{out}
GroupNorm	–	–	C_{out}
Skip path	$1 \times 1 \times 1$	s	C_{out}
ReLU	–	–	C_{out}

Figure 5.8: Schematic illustration of a residual block used in the modality-specific expert networks. Each block contains two 3D convolutional layers with group normalization and nonlinear activation. The table on the right describes the parameters of each layer.

In addition to the convolutional operations, the residual blocks include a lightweight spatial attention mechanism inspired by [42], illustrated in Figure 5.9. This module is placed inside each residual block, after the last group normalization but before adding the residual connection. It computes an attention mask from channel-wise average and maximum responses and uses it to reweight the volumetric feature tensor. The purpose is to allow the encoder to emphasize spatial regions that are more informative for diagnosis, while suppressing less relevant responses.

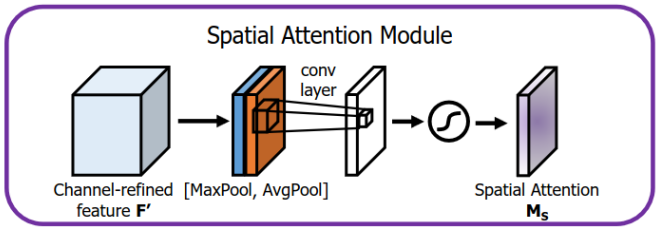
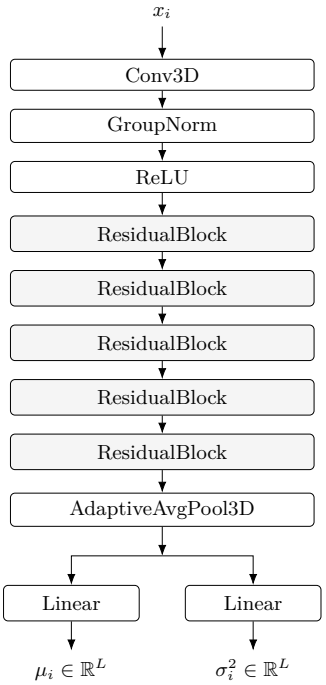


Figure 5.9: Spatial Attention Module from CBAM [42]. $\mathbf{F}'$ represents the features in input to the layer.

The modality encoders are ultimately composed of a series of stacked residual blocks. An initial convolution, normalization and activation layers precede the stack, which is followed at the end by an average pooling layer and two parallel linear layers, which produce the parameters μ_i and the log-variances vector $\log \sigma_i^2$. The use of log-variances improves numerical stability and ensures that the variances σ_i^2 are positive after exponentiation. The high-level composition of the network is shown in Figure 5.10.

Layer	Parameters
Linear	$L \mapsto 128$
ReLU	-
Linear	$128 \mapsto 3$

Table 5.4: Layer and parameters of the shared classifier.



Layer	Parameters
Conv3D	kernel $k_0 = 3$, stride $s_0 = 1$, channels $C_0 = 8$
GroupNorm	-
ReLU	-
ResidualBlock 1	channels $C_1 = 8$, stride $s_1 = 1$
ResidualBlock 2	channels $C_2 = 16$, stride $s_2 = 2$
ResidualBlock 3	channels $C_3 = 16$, stride $s_3 = 1$
ResidualBlock 4	channels $C_4 = 32$, stride $s_4 = 2$
ResidualBlock 5	channels $C_5 = 32$, stride $s_5 = 1$
AdaptiveAvgPool3D	output size $1 \times 1 \times 1$
Linear	$C_5 \mapsto L$, output μ_i
Linear	$C_5 \mapsto L$, output σ_i^2

Figure 5.10: Schematic illustration of a modality encoder network. The input modality x_i is processed by an initial convolutional block followed by five residual blocks and adaptive average pooling. The resulting representation is passed to two parallel linear layers that output the parameters μ_i and σ_i^2 of the auxiliary Gaussian expert. In the table on the right, C_i and s_i denote the output channels and the stride of each layer, where applicable.

It is important to remark that MRI and FDG-PET encoders have the same architecture but do not share parameters. This choice ensures that both modalities are mapped to the same type of latent representation, while still allowing each encoder to learn modality-specific features.

5.2.2. Shared classifier

Let $\mathcal{P}(\{0,1,2\})$ denote the space of discrete probability distributions over $\{0,1,2\}$. The shared classifier $\hat{p}_\theta(y|\mathbf{z}) : \mathbb{R}^L \to \mathcal{P}(\{0,1,2\})$ maps latent representations $\mathbf{z}$ to a probability distribution over the diagnostic classes CN, MCI and AD, corresponding to the integers 0, 1 and 2. The classifier is parametrized by a small feedforward network composed of one hidden layer, a ReLU activation, and a final linear layer producing three logits corresponding to the CN, MCI and AD classes. The class probabilities are obtained by applying the softmax function to these logits. Since the same classifier is used for all modality availability patterns, the model is encouraged to learn a common diagnostic decision space shared by MRI-only, FDG-PET-only and multimodal inputs. The parameters of the layers are indicated in Table 5.4.

The deployed Product-of-Experts model, comprising the two modality-specific encoders and the shared classifier, contains 334,219 trainable parameters. The meta-learned regularizer parameters ψ and the domain-specific classifier copies used in Algorithm 1 are not included in this count, since they enter only the training objective and are discarded once the regularizer has been

learned.

For reference, the number of trainable parameters of DiaMond is not reported in [1]. Instantiating the publicly available implementation with the configuration adopted in that work yields 30,106,015 trainable parameters, making the proposed model approximately 90 times smaller. This count depends on the width of the feed-forward network, which is not specified in [1] and is taken here from the released implementation.

5.3. Stochastic and Deterministic objective

As described in the previous chapters, the predictive distributions $\hat{p}_{\theta,\phi_i}(y|x_i)$ and $\hat{p}_{\theta,\phi}(y|x_1, x_2)$ are expressed as expectations with respect to the latent variable $\mathbf{z}$. In the present model, these expectations do not have a closed form solution, mainly because the classifier $\hat{p}_\theta(y|\mathbf{z})$ is parametrized by a neural network. It is therefore necessary to approximate them. In this section, we describe two possible approximations, one deterministic and one stochastic, and discuss their respective advantages.

Deterministic mean approximation

The first approach approximates the expectation $\mathbb{E}_\mathbf{z}[\hat{p}_\theta(y|\mathbf{z})]$ by evaluating the classifier at the expected value of the latent variable, namely $\hat{p}_\theta(y|\mathbb{E}[\mathbf{z}])$. This approximation can be motivated by a first order Taylor expansion. For a fixed class y, let $f(\mathbf{z}) = \hat{p}_\theta(y|\mathbf{z})$, and let $\mathbf{z}^* = \mathbb{E}[\mathbf{z}]$. When $\mathbf{z}$ is sufficiently close to its expected value $\mathbf{z}^*$, i.e. when the variance of the latent variable is sufficiently small, a first order Taylor approximation of f around $\mathbf{z}^*$ gives

$$f(\mathbf{z}) \approx f(\mathbf{z}^*) + \nabla_\mathbf{z} f(\mathbf{z}^*)^T(\mathbf{z} - \mathbf{z}^*).$$

Taking expectations on both sides yields

$$\mathbb{E}[f(\mathbf{z})] \approx f(\mathbf{z}^*) + \nabla_\mathbf{z} f(\mathbf{z}^*)^T(\mathbb{E}[\mathbf{z}] - \mathbf{z}^*) = f(\mathbf{z}^*).$$

Therefore, substituting back $f(\mathbf{z}) = \hat{p}_\theta(y|\mathbf{z})$, we obtain

$$\mathbb{E}_\mathbf{z}[\hat{p}_\theta(y|\mathbf{z})] \approx \hat{p}_\theta(y|\mathbf{z} = \mathbb{E}[\mathbf{z}]).$$

This approximation is deterministic and easy to implement, since the full predictive model becomes a composition of differentiable functions. It is therefore directly compatible with gradient based optimization. However, it relies on two conditions that are difficult to verify in practice. First, the classifier should be close to linear in the region where $\mathbf{z}$ is likely to fall, so that the second and higher order terms of the expansion can be neglected. Since the classifier is a neural network, little can be said about its curvature, and this condition cannot be checked directly.

Second, the variances produced by the encoders should be small, so that $\mathbf{z}$ stays close to its mean. A partial safeguard comes from the prior. By Proposition 3.4.1, the covariance of the latent conditional is $\Sigma_S = (I + \sum_{i \in S} \Sigma_i^{-1})^{-1}$, whose precision is strictly larger than that of the prior. The variances of $\mathbf{z}$ are therefore always smaller than one, regardless of what the encoders output, and they decrease further when more modalities are available. This bounds the spread of the latent variable, although it does not by itself guarantee that the approximation is accurate.

Monte Carlo approximation

A second approach is to approximate the expectation using a Monte Carlo estimator. If $\mathbf{z}_i \overset{\text{i.i.d.}}{\sim} \hat{p}_\phi(\mathbf{z}|x_1, x_2)$, then the Law of Large Numbers gives

$$\frac{1}{S}\sum_{i=1}^{S}\hat{p}_\theta(y|\mathbf{z}_i) \xrightarrow[S\to\infty]{a.s.} \mathbb{E}_{\mathbf{z}\sim\hat{p}_\phi}[\hat{p}_\theta(y|\mathbf{z})].$$

This approximation is more faithful to the original predictive distribution, since it averages the classifier over samples drawn from the latent distribution.

However, sampling introduces a difficulty for gradient based optimization. If the samples $\mathbf{z}_i$ are treated as direct realizations from $\hat{p}_\phi(\mathbf{z}|x_1, x_2)$, then they do not provide a differentiable path with respect to ϕ. One possibility would be to estimate the gradient directly. Under suitable regularity assumptions, one obtains

$$\nabla_\phi \mathbb{E}_{\mathbf{z}\sim\hat{p}_\phi}[\hat{p}_\theta(y|\mathbf{z})] = \mathbb{E}_{\mathbf{z}\sim\hat{p}_\phi}[\nabla_\phi \log \hat{p}_\phi(\mathbf{z}|x_1, x_2)\hat{p}_\theta(y|\mathbf{z})].$$

This estimator can be approximated by Monte Carlo sampling, but it is known to suffer from high variance. A lower variance alternative is given by the reparametrization trick [43]. The idea is to express $\mathbf{z}$ as a deterministic function of the parameters ϕ and of an auxiliary random variable $\boldsymbol{\varepsilon}$ whose distribution does not depend on ϕ. In the Gaussian case considered here, where $\mathbf{z} \sim \mathcal{N}(\mu_\phi, \mathrm{diag}(\sigma_\phi^2))$, we can write $\mathbf{z} = \mu_\phi + \sigma_\phi \circ \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ and where $\circ$ denotes element wise multiplication. The expectation can then be rewritten as

$$\mathbb{E}_{\mathbf{z}\sim\hat{p}_\phi}[\hat{p}_\theta(y|\mathbf{z})] = \mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{p}_\theta(y|\mu_\phi + \sigma_\phi \circ \boldsymbol{\varepsilon})].$$

The resulting Monte Carlo estimator is

$$\frac{1}{S}\sum_{i=1}^{S}\hat{p}_\theta(y|\mu_\phi + \sigma_\phi \circ \boldsymbol{\varepsilon}_i), \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, I).$$

Since the randomness is now isolated in $\boldsymbol{\varepsilon}_i$, whose distribution is independent of the model parameters, the estimator can be differentiated with respect to ϕ by standard backpropagation.

In practical implementations, the cross-entropy loss function usually accepts directly the logits from the network in order to avoid numerical stability issues. In our case this would represent a problem, since the stochastic approximation is the average over probabilities, rather than the average over logits. To overcome this issue, in the implementation we develop a custom version of the cross-entropy loss that accepts S logits triples deriving from the MC samples, and then computes the cross-entropy loss while avoiding numerical instabilities.

5.4. Evaluation Metrics

All the experiments share the common three-class CN vs MCI vs AD classification task. As discussed in Section 5.1.6, these classes are imbalanced, therefore it is important to evaluate the performance of the model in a way that accounts for this imbalance, as otherwise standard metrics may mask poor performance on minority classes.

A natural first choice would be overall accuracy, defined as the fraction of correctly classified samples. However, accuracy is dominated by the most frequent class, MCI in our case, and can therefore remain high even when the model performs poorly on CN or AD. Since correctly identifying all three diagnostic categories is equally important from a clinical perspective, a metric that weighs each class equally, regardless of its frequency in the dataset, is preferable. For this reason, we adopt balanced accuracy and macro-averaged F1 score as the main evaluation metrics throughout this thesis, following common practice in the ADNI classification literature [1].

Let $\mathcal{Y} = \{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$ denote the set of classes, and for each class $c \in \mathcal{Y}$, let TP_c, FN_c and FP_c denote respectively the number of true positives, false negatives and false positives for class c, obtained by treating c as the positive class and all other classes as negative.

Definition 5.4.1 (Balanced accuracy)**.** The recall of class $c \in \mathcal{Y}$ is

$$\text{Recall}_c := \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}.$$

The balanced accuracy (BACC) is the unweighted average of the per-class recall,

$$\text{BACC} := \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{Recall}_c.$$

Definition 5.4.2 (Macro F1 score)**.** The precision of class $c \in \mathcal{Y}$ is

$$\text{Precision}_c := \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c},$$

and its F1 score is the harmonic mean of precision and recall,

$$\text{F1}_c := \frac{2\,\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

The macro F1 score is the unweighted average of the per-class F1 scores,

$$\text{Macro F1} := \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{F1}_c.$$

By averaging across classes rather than across samples, both metrics assign equal importance to CN, MCI and AD regardless of their frequency in the test set. Balanced accuracy focuses exclusively on how well each true class is recovered, while macro F1 also penalizes over-prediction of a given class through the precision term. The two metrics are therefore complementary.

As a reference point for the difficulty of this task, [1] recently reported a balanced accuracy of 65.2% and a macro F1 score of 64.9% on the CN vs MCI vs AD classification task using ADNI data. Notably, the same study reports substantially higher performance on the binary CN vs AD classification task, with a balanced accuracy of 92.42%. This gap between the two- and three-class settings is consistent with the discussion in Section 2.4, and further confirms that the added difficulty of the three-class problem lies primarily in the MCI class, which is clinically heterogeneous and shares overlapping characteristics with both CN and AD, making it substantially harder to recognize than the two more clearly separated diagnostic categories.

5.5. Experiment design

The experiments are designed to evaluate whether the proposed Product-of-Experts model can exploit incomplete multimodal data while remaining reliable when one modality is missing. The main classification task is the three-class diagnosis problem with labels CN, MCI and AD. Since the class distribution is not perfectly balanced, and since MCI is expected to be the most difficult class, performance is reported using balanced accuracy (BACC) and macro F1-score, as defined in the previous section. Confusion matrices are also used to inspect the class-wise behaviour of the models.

All final model comparisons are evaluated on a fixed held-out test set. The remaining data are partitioned into five folds, as described in Section 5.1.7. For each experiment, five models are trained using a standard cross-validation procedure: at each run, four folds are used for training and the remaining fold is used for validation. Thus, each experiment yields five trained models, one per fold.

Several of the experiments described below require choosing between competing design options, namely the choice between the deterministic and stochastic predictive approximations, and the choice of the regularization strength λ_{KL}. Both choices are made using the same model selection procedure, described next, before any model is evaluated on the held-out test set.

5.5.1. Model selection procedure

Several experiments below require choosing between competing configurations, such as the deterministic versus stochastic predictive approximation, or different values of the regularization strength λ_{KL}. In all cases, the same procedure is used: for each candidate configuration, a model is trained separately on each of the five folds, and for each fold the epoch achieving the highest validation BACC is retained. These five best-epoch values are then averaged across folds, giving one fold-averaged validation BACC per configuration. The configuration with the highest fold-averaged validation BACC is selected, and only the five models trained under this configuration are evaluated on the held-out test set.

5.5.2. Stochastic vs Deterministic approximation and Baseline experiments

The first set of experiments determines which approximation of the predictive distribution performs best on the present task: the deterministic mean approximation or the stochastic Monte Carlo approximation. At the same time, these experiments establish baseline performances for simple unimodal and paired multimodal models.

For each approximation, we train three baselines: an MRI expert on all MRI-available samples, an FDG-PET expert on all FDG-PET-available samples, and a paired multimodal baseline, using the same Product-of-Experts fusion rule but restricted to samples for which both MRI and FDG-PET are available. Let D_1, D_2 and $D_{1,2}$ denote the MRI-available set, the FDG-PET-available set and the complete MRI–FDG-PET set, respectively, as defined in Definition 3.1.2, and let $\mathcal{L}_{\mathrm{CE}}$ denote the cross-entropy loss. The three baselines are trained using the objectives

$$\mathcal{L}_{\mathrm{MRI}} := \frac{1}{|D_1|} \sum_{(x_1,y)\in D_1} \mathcal{L}_{\mathrm{CE}}\big(\hat{p}_{\theta,\phi_1}(\cdot|x_1), y\big),$$

$$\mathcal{L}_{\text{FDG-PET}} := \frac{1}{|D_2|} \sum_{(x_2,y)\in D_2} \mathcal{L}_{\mathrm{CE}}\big(\hat{p}_{\theta,\phi_2}(\cdot|x_2), y\big),$$

$$\mathcal{L}_{\text{MRI+FDG-PET}} := \frac{1}{|D_{1,2}|} \sum_{(x_1,x_2,y)\in D_{1,2}} \mathcal{L}_{\mathrm{CE}}\big(\hat{p}_{\theta,\phi}(\cdot|x_1,x_2), y\big).$$

For each of these three baselines, the deterministic and stochastic approximations are compared using the model selection procedure described in Section 5.5.1. In our experiments, the deterministic approximation achieves the higher fold-averaged validation BACC for all three baselines, and is therefore selected for all subsequent experiments. The five deterministic models corresponding to each baseline are then evaluated on the held-out test set.

Once the predictive approximation has been fixed, we train a basic version of the proposed Product-of-Experts model on all three modality-availability sets D_1, D_2 and $D_{1,2}$ jointly, using the combined supervised loss

$$\mathcal{L}_{\mathrm{sup}} := \mathcal{L}_{\mathrm{MRI}} + \mathcal{L}_{\text{FDG-PET}} + \mathcal{L}_{\text{MRI+FDG-PET}},$$

and evaluate the resulting five models on the held-out test set.

5.5.3. Mini-batch construction for the joint objective

Due to the high dimensionality of each neuroimaging acquisition and the large number of parameters in the model, gradients cannot be computed on entire loss terms directly, and mini-batches of size 8 are used instead. For the single-set baselines, trained on either D_1, D_2 or $D_{1,2}$ alone, this is straightforward. Training the full Product-of-Experts model on $\mathcal{L}_{\mathrm{sup}}$, however, requires combining gradients from D_1, D_2 and $D_{1,2}$ simultaneously, which introduces an additional design choice.

At each gradient update, one batch of 8 samples is drawn from each of D_1, D_2 and $D_{1,2}$, so that all three loss terms contribute equally to every update. Since D_1 is roughly four times larger than D_2 and $D_{1,2}$, matching the number of batches per epoch across the three sets would require either enlarging the D_1 batch size proportionally, which is computationally too expensive, or shrinking the batches for D_2 and $D_{1,2}$, which would make their gradient estimates excessively noisy. We therefore keep the batch size fixed at 8 for all three sets, and define one epoch as a full pass over the smallest set, $D_{1,2}$, while D_1 and D_2 are only partially traversed within that epoch. The data loaders for all three sets are reshuffled at every epoch, so that different subsets of D_1 and D_2 are seen across epochs, and the full sets are eventually covered over the course of training. This keeps training computationally feasible, but has the side effect that D_1 is visited less often per epoch than D_2 and $D_{1,2}$, which are of comparable size. As discussed in Chapters 6 and 7, this may affect the performance of the MRI prediction pathway relative to the other two.

5.5.4. MetaReg and KL regularization experiments

The second set of experiments evaluates the additional training objectives introduced in Chapter 4. Starting from the basic Product-of-Experts model, we compare three variants, PoE+KL, PoE+MetaReg, and PoE+KL+MetaReg, obtained by adding the KL-alignment term and the meta-learned classifier regularizer, defined in (4.3) and (4.1) respectively, to the supervised objective:

$$\mathcal{L} := \mathcal{L}_{\text{sup}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}} + \lambda_R\,R_\psi(\theta).$$

Since each Product-of-Experts model can be evaluated under three inference settings, namely MRI-available, FDG-PET-available, and complete MRI+FDG-PET inference, the selection criterion for λ_{KL} cannot rely on a single validation BACC as in the previous experiments. Instead, we define the per-epoch selection metric as the arithmetic mean of the validation BACC across the three inference settings,

$$\text{BACC}^{(t)}_{\text{avg}} := \frac{1}{3}\left(\text{BACC}^{(t)}_{\text{MRI}} + \text{BACC}^{(t)}_{\text{FDG-PET}} + \text{BACC}^{(t)}_{\text{MRI+FDG-PET}}\right). \tag{5.1}$$

where t denotes the epoch. The PoE+KL model is trained separately for four candidate values of λ_{KL}, with $\lambda_R = 0$. The best value $\lambda^\star_{\text{KL}}$ is then selected using the same model selection procedure described in Section 5.5.1, using BACC_{avg} in place of the validation BACC. The model trained with the selected value $\lambda^\star_{\text{KL}}$ is then evaluated on the held-out test set.

The MetaReg regularizer is instead enabled or disabled by setting $\lambda_R = 1$ or $\lambda_R = 0$, respectively. Since the weights ψ of the MetaReg regularizer are themselves meta-learned during training, no additional sweep over λ_R is performed, and $\lambda_R = 1$ is used whenever MetaReg is enabled. For the PoE+KL+MetaReg experiment, we use the value $\lambda^\star_{\text{KL}}$ selected from the PoE+KL sweep above, combined with $\lambda_R = 1$. The hyperparameter values used for each variant are summarized in Table 5.5.

Table 5.5: Regularization hyperparameters used in the MetaReg and KL experiments. The basic PoE model corresponds to $\lambda_{\text{KL}} = \lambda_R = 0$.

Variant	λ_{KL}	λ_R
PoE + KL	$2 \cdot 10^{-4}$	0
	$1 \cdot 10^{-3}$	0
	$5 \cdot 10^{-3}$	0
	$1 \cdot 10^{-2}$	0
PoE + MetaReg	0	1
PoE + KL + MetaReg	$1 \cdot 10^{-2}$	1

For complete test samples (x_1, x_2, y), each Product-of-Experts model is evaluated under the same three inference settings used for model selection: an MRI-available setting, in which prediction

relies only on the MRI expert; an FDG-PET-available setting, in which prediction relies only on the FDG-PET expert; and a multimodal setting, in which both experts are combined through the Product-of-Experts rule. This allows us to evaluate whether the same trained model remains useful when one modality is absent at inference time.

Most architectural and optimization hyperparameters are kept fixed across experiments, based on preliminary runs and computational feasibility. In particular, the network architecture, learning rate, weight decay, dropout, batch size and latent dimension are not re-tuned separately for every model.

6

Results

This chapter reports the experimental results of this thesis. Section 6.1 compares the deterministic and stochastic predictive approximations and establishes the unimodal and multimodal baselines. Section 6.2 evaluates the basic Product-of-Experts model against these baselines. Sections 6.3 to 6.5 then examine the effect of KL regularization, MetaReg, and their combination, respectively, on the performance of the Product-of-Experts model.

6.1. Stochastic vs Deterministic approximations and Baseline experiments

Table 6.1: Fold-averaged validation balanced accuracy for the deterministic and stochastic predictive approximations, for each of the three baselines. Values are used to select the approximation adopted in all subsequent experiments, following the model selection procedure of Section 5.5.1. The best value per row is shown in bold.

Baseline	Deterministic	Stochastic
MRI	**0.6053 ± 0.0231**	0.5664 ± 0.0333
FDG-PET	**0.6256 ± 0.0309**	0.4439 ± 0.1015
MRI + FDG-PET	**0.6079 ± 0.0348**	0.4152 ± 0.1024

Table 6.2: Test-set performance of the deterministic MRI, FDG-PET and MRI+FDG-PET baselines, trained on D_1, D_2 and $D_{1,2}$ respectively. Results are reported as mean ± standard deviation across the five (one per fold) evaluations on the test set.

Baseline	Balanced accuracy	Macro F1-score
MRI	0.5835 ± 0.0118	0.5324 ± 0.0052
FDG-PET	0.6411 ± 0.0291	0.5656 ± 0.0344
MRI + FDG-PET	0.6566 ± 0.0389	0.5490 ± 0.0429

Table 6.1 reports the fold-averaged validation balanced accuracy for the deterministic and stochastic approximations across the three baselines. In all three cases, the deterministic approximation achieves a higher validation BACC, and is therefore selected for all subsequent experiments, following the model selection procedure described in Section 5.5.1. The size of this gap, however, is not uniform: for the MRI baseline the two approximations are relatively close (0.6053 vs. 0.5664), whereas for FDG-PET and MRI+FDG-PET the gap is considerably larger (0.6256 vs. 0.4439, and 0.6079 vs. 0.4152, respectively), and the stochastic approximation also shows a substantially higher standard deviation in these two settings.

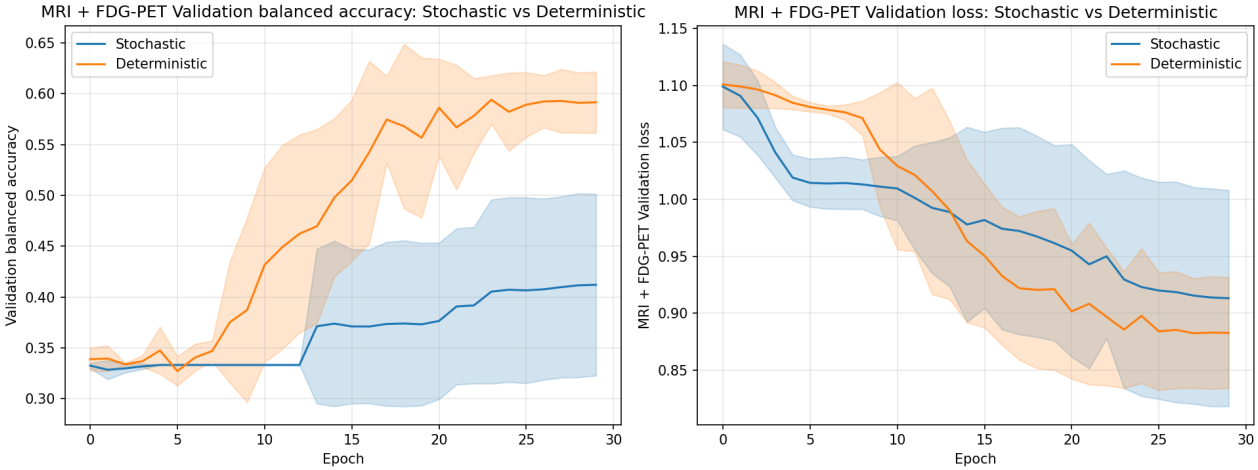


Figure 6.1: Comparison of validation BACC and validation loss in the MRI+FDG–PET setting for each training epoch, averaged over the folds. The shaded areas indicate a ± 1 standard deviation range.

To better understand this pattern, we inspect the fold-averaged validation BACC and validation loss curves during training (Figures 6.1 and 6.3) for the MRI and MRI + FDG-PET cases, together with the respective confusion matrices computed on the held-out test set for the selected checkpoint of each approximation. Since per-epoch validation confusion matrices were not logged during training, these matrices are used here purely as a diagnostic tool to interpret the observed validation gap, and play no role in the model selection, which relies exclusively on the validation BACC values reported in Table 6.1. The multimodal MRI + FDG-PET is selected because it features the largest BACC gap between deterministic and stochastic approximation, while the MRI setting the lowest; the FDG-PET case is omitted as it is qualitatively similar to the multimodal setting.

In the MRI+FDG-PET setting (Figure 6.1), the deterministic approximation reaches a substantially higher validation BACC than the stochastic one, while the two validation loss curves remain comparable throughout training. Such a discrepancy may arise because cross-entropy loss and balanced accuracy measure different aspects of the predicted probability distributions. Cross-entropy rewards high confidence on the true class, whereas balanced accuracy is exclusively based on final predictions that depend only on the ordering of these probabilities. Therefore, a probability distribution with low CE loss may still be producing a wrong prediction, while a probability distribution with low confidence on the true class, yielding a high CE loss, may still assign the label correctly.

A possible explanation specific to our case is that the stochastic objective may favour probability distributions that reduce average cross-entropy without improving the ordering of the class probabilities. This effect is particularly relevant for CN and MCI, which are harder to separate than AD. In ambiguous cases, a model may assign similar probabilities to CN and MCI. For example, a CN subject predicted as (0.41,0.39,0.20) and an MCI subject predicted as (0.39,0.41,0.20) would both be classified correctly, although with relatively high cross-entropy, since the model is uncertain. Conversely, a model biased toward MCI could output probabilities closer to (0.40,0.60,0.00) for both subjects. This would reduce the loss on MCI samples, since $-\log 0.60 < -\log 0.41$, while leaving the loss on CN samples nearly unchanged, since $-\log(0.40) \approx -\log(0.41)$. As a result, the average cross-entropy may decrease even though CN samples are misclassified more often.

This interpretation is consistent with the confusion matrices in Figure 6.2: the stochastic model classifies 100% of CN samples and 67.4% of AD samples as MCI, while correctly predicting MCI in 91.9% of cases, suggesting that its predicted probabilities are systematically shifted toward MCI rather than genuinely uncertain between classes. In addition, the validation loss of the stochastic approximation does not show a clear plateau by the end of training, unlike the deterministic one, suggesting that the stochastic model may not fully converge within the given training budget.

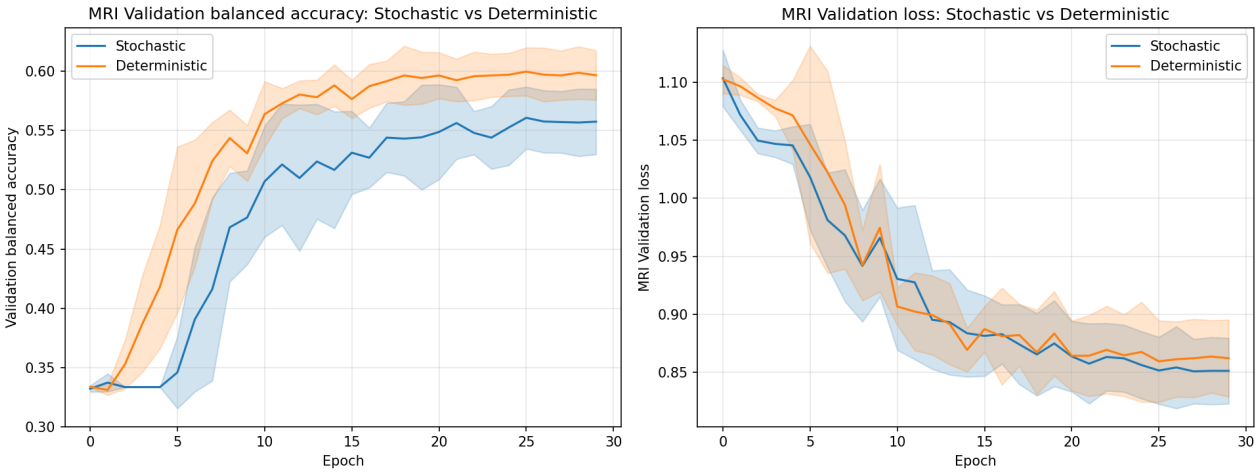


Figure 6.3: Comparison of validation BACC and validation loss in the MRI setting for each training epoch, averaged over the folds. The shaded areas indicate a ± 1 standard deviation range.

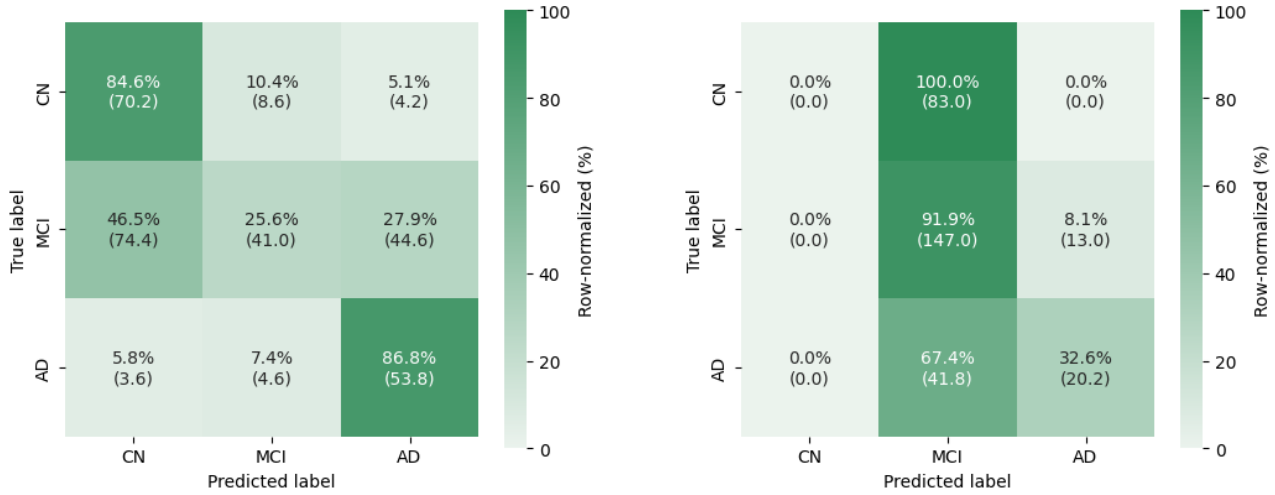


Figure 6.2: Test-set average row-normalized confusion matrices across folds for the deterministic (left) and stochastic (right) objectives in the MRI+FDG-PET setting. Rows denote true labels and columns denote predicted labels. In each cell, the average number of samples is indicated between parentheses.

The MRI setting shows a different pattern (Figure 6.3). Here, the validation BACC curves of the two approximations are closer together, and the shaded standard deviation bands are visibly narrower than in the MRI+FDG-PET setting. This is also reflected in the confusion matrices (Figure 6.4): unlike the MRI+FDG-PET case, the stochastic MRI model does not collapse toward predicting a single class. It correctly classifies CN in 52.8% of cases and AD in 49.8% of cases, considerably lower than the deterministic model, but far from the near-total MCI collapse observed in the MRI+FDG-PET setting. The FDG-PET-only setting is omitted here, as its validation curves and confusion matrices follow a pattern very similar to the MRI+FDG-PET case, consistent with the large validation gap and high standard deviation already visible in Table 6.1.

Taken together, these results suggest that the instability of the stochastic approximation may be related to the amount of available training data. The MRI baseline, trained on the largest set D_1, shows a comparatively mild degradation relative to the deterministic approximation, whereas the FDG-PET and MRI+FDG-PET baselines, trained on the considerably smaller sets D_2 and $D_{1,2}$,

show a much larger performance gap and a pronounced tendency to collapse toward the MCI class. We note that this observation is based on only three settings that differ in more than just sample size, such as modality and class balance, and should therefore be treated as a hypothesis rather than an established explanation.

Overall, the deterministic approximation consistently outperforms the stochastic one in terms of validation balanced accuracy, across all three baselines. For this reason, the deterministic approximation is adopted for all subsequent experiments in this thesis.

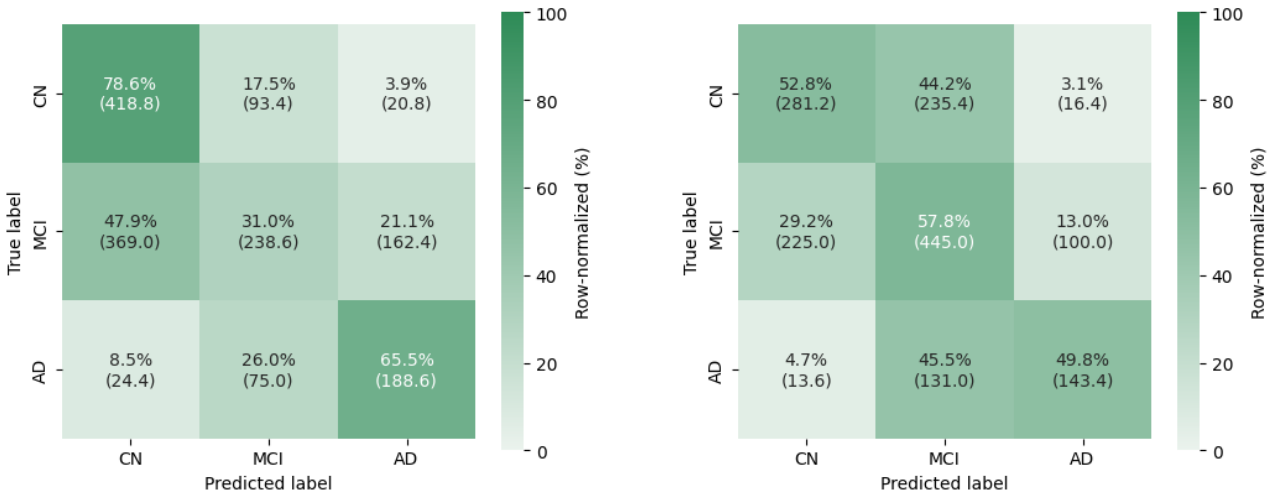


Figure 6.4: Test-set average row-normalized confusion matrices across folds for the deterministic (left) and stochastic (right) objectives in the MRI setting. Rows denote true labels and columns denote predicted labels. In each cell, the average number of samples is indicated between parentheses.

6.2. Basic Product-of-Experts

Table 6.3: Test-set comparison between the deterministic single-pathway baselines and the basic Product-of-Experts model, trained jointly on D_1, D_2 and $D_{1,2}$. All methods are evaluated under the same test-time modality availability. Results are reported as mean ± standard deviation across the five (one per fold) evaluations on the test set. The best value per test-time input and metric is shown in bold.

Test-time input	Model	Balanced accuracy	Macro F1-score
MRI-available	MRI baseline	**0.5835 ± 0.0118**	**0.5324 ± 0.0052**
	PoE, MRI inference	0.5118 ± 0.0286	0.4374 ± 0.0192
FDG-PET-available	FDG-PET baseline	**0.6411 ± 0.0291**	**0.5656 ± 0.0344**
	PoE, FDG-PET inference	0.6180 ± 0.0217	0.5319 ± 0.0180
MRI + FDG-PET	MRI + FDG-PET baseline	**0.6566 ± 0.0389**	**0.5490 ± 0.0429**
	PoE, multimodal inference	0.6427 ± 0.0329	0.5476 ± 0.0291

Table 6.3 compares the deterministic single-pathway baselines to the basic Product-of-Experts model, trained jointly on D_1, D_2 and $D_{1,2}$, under matching test-time modality availability. Across all three inference settings, the corresponding baseline outperforms the basic PoE model, both in terms of balanced accuracy and macro F1-score. This indicates that, in its basic form, jointly training the Product-of-Experts model on all modality-availability sets using a standard supervised objective does not improve performance relative to simply training a dedicated model for each modality configuration. This finding is consistent with the motivation given in Chapter 4 for the proposed regularization techniques, namely that the latent representations induced by different

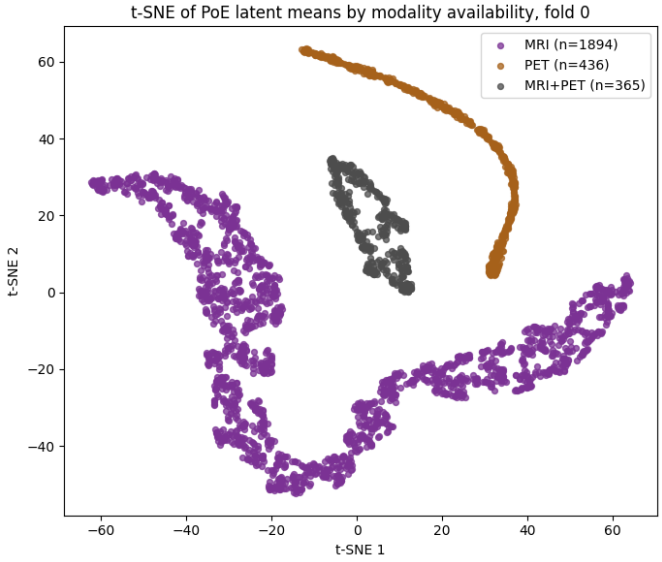


Figure 6.5: t-SNE plot of the latent embeddings of the basic Product-of-Expert model trained earlier, produced using the samples from the first cross-validation fold. The colors indicate the different modality-availability patterns that produced the embeddings.

modality-availability patterns may act as distinct input regimes for the shared classifier, and that naive joint optimization across these regimes may be suboptimal. However, this result does not by itself demonstrate that the proposed regularization techniques are necessary or effective; it only shows that the underlying concern motivating their introduction is supported experimentally, at least for the basic PoE model considered here.

The size of this performance gap, however, is not uniform across the three settings. It is considerably larger for the MRI pathway (7.2 points of balanced accuracy, 9.5 points of macro F1) than for the FDG-PET pathway (2.3 and 3.4 points, respectively) or the multimodal pathway (1.4 points of balanced accuracy, and an essentially negligible 0.1 points of macro F1). A possible explanation for this asymmetry lies in the mini-batch construction procedure described in Section 5.5.3. At each gradient update, one batch of equal size is drawn from each of D_1, D_2 and $D_{1,2}$, so the three pathways receive the same number of gradient updates over training. However, since one epoch is defined as a full pass over the smallest set $D_{1,2}$, and D_1 is roughly four to five times larger than D_2 and $D_{1,2}$, each epoch only covers a fraction of D_1, while D_2 and $D_{1,2}$ are traversed (almost) fully. Although the data loaders are reshuffled at every epoch, so that all of D_1 is eventually used over the course of training, each individual MRI sample is on average seen fewer times across the fixed training budget than each FDG-PET or complete sample. This may explain why the MRI pathway is the most affected by the transition from individual baseline to joint training: although the MRI expert and the shared classifier receive as many gradient updates from MRI data as from the other two modalities, the effective exposure to the full diversity of D_1 over training is reduced relative to the dedicated MRI baseline, which is trained on the entirety of D_1 at every epoch.

To further investigate why joint training does not improve, and in the case of MRI substantially harms, classification performance, Figure 6.5 shows a two-dimensional t-SNE projection of the latent means produced by the basic PoE model on the validation data of the first fold, taken as representative. t-SNE (t-distributed Stochastic Neighbour Embedding) [44] is a machine learning algorithm used to reduce the dimensionality of a dataset. It begins by first defining a similarity distance for each pair of original datapoints, and constructs a probability distribution p assigning higher probability to points that are more similar to each other. Secondly, for each original datapoint, a point in the reduced-dimension space is created, and an analogous probability distribution q is constructed, such that pairs of points that are closer have higher probability. Then, the KL divergence between p and q is minimized by moving the lower dimensional points

using gradient descent.

In the t-SNE plot in Figure 6.5, the MRI-only, FDG-PET-only and multimodal encodings form three visually distinct clusters, with clear separation and no overlap between them. The plots of the remaining folds are omitted but are qualitatively equivalent to that reported in 6.5 This suggests that, in the absence of an explicit alignment mechanism, the shared classifier $\hat{p}_\theta(y|\mathbf{z})$ receives latent inputs whose distribution depends strongly on which modalities were available for a given sample. Since the same classifier parameters are used across all three regimes, a decision boundary that is well suited to one region of the latent space may be poorly suited to another, providing a plausible mechanism for the performance gap observed in Table 6.3. This observation directly motivates the KL-based alignment term and the meta-learned classifier regularizer introduced in Chapter 4, both of which are evaluated in the following sections.

6.3. Product-of-Experts with KL regularization

As described in Section 5.5.4, the KL regularization strength λ_{KL} is selected by comparing four candidate values using the fold-averaged validation metric BACC_{avg}, defined as the arithmetic mean of the validation balanced accuracy across the MRI, FDG-PET and multimodal inference pathways (See Eq. 5.1). Table 6.4 reports this metric for each candidate value. The value $\lambda^\star_{\text{KL}} = 10^{-2}$ achieves the highest BACC_{avg} and is therefore selected for all subsequent KL-regularized experiments.

Table 6.4: Fold-averaged validation BACC_{avg} (Equation 5.1) for each candidate value of λ_{KL}, used to select $\lambda^\star_{\text{KL}}$ following the model selection procedure of Section 5.5.1. The selected value is shown in bold.

λ_{KL}	BACC_{avg}
$2 \cdot 10^{-4}$	0.6120 ± 0.0196
$1 \cdot 10^{-3}$	0.6195 ± 0.0212
$5 \cdot 10^{-3}$	0.6159 ± 0.0184
$1 \cdot 10^{-2}$	**0.6200 ± 0.0219**

Table 6.5: Test-set balanced accuracy and macro F1-score of the Product-of-Experts model with KL regularization ($\lambda^\star_{\text{KL}}$, selected in Table 6.4), compared to the basic PoE model ($\lambda_{\text{KL}} = 0$), across the three inference pathways. Results are reported as mean ± standard deviation across the five (one per fold) evaluations on the test set. The best value per column is shown in bold.

Model	MRI pathway	FDG-PET pathway	MRI + FDG-PET pathway
Balanced accuracy			
PoE basic	0.5118 ± 0.0286	**0.6180 ± 0.0217**	0.6427 ± 0.0329
PoE + KL	**0.5553 ± 0.0238**	0.5923 ± 0.0393	**0.6529 ± 0.0169**
Macro F1-score			
PoE basic	0.4374 ± 0.0192	**0.5319 ± 0.0180**	0.5476 ± 0.0291
PoE + KL	**0.4994 ± 0.0180**	0.5264 ± 0.0279	**0.5635 ± 0.0117**

Table 6.5 compares the test-set performance of the basic PoE model and the KL-regularized model ($\lambda^\star_{\text{KL}} = 10^{-2}$) across the three inference pathways. KL regularization improves the MRI pathway considerably, by 4.4 points of balanced accuracy and 6.2 points of macro F1-score. The multimodal pathway also improves, though to a smaller extent, gaining 1.0 point of balanced accuracy and 1.6 points of macro F1-score. In contrast, the FDG-PET pathway shows a moderate decline of 2.6 points in balanced accuracy and a slight decline of 0.6 points in macro F1-score.

Comparing these results to the original baselines (Table 6.2), KL regularization narrows, but does not close, the gap for the MRI and FDG-PET pathways. The MRI pathway remains 2.8 points of balanced accuracy and 3.3 points of macro F1-score below the MRI baseline, while the

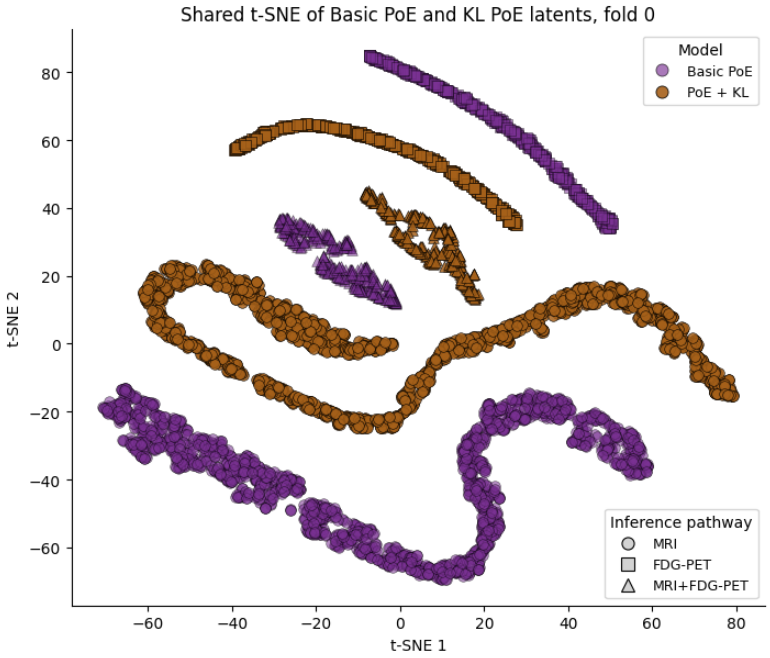


Figure 6.6: t-SNE plot of the embeddings derived from the basic PoE and the KL-regularized PoE. Here, the two colors indicate the different models producing the embeddings, namely the basic PoE and the KL-regularized PoE. The shape of marker indicates instead the diagnostic category. The embeddings are computed from the data of the first cross-validation fold.

FDG-PET pathway remains approximately 5 points of balanced accuracy and 4 points of macro F1-score below the FDG-PET baseline. The multimodal pathway, however, benefits the most from regularization relative to the baseline: its balanced accuracy is within 0.4 points of the MRI+FDG-PET baseline, and its macro F1-score exceeds the baseline by 1.5 points.

Table 6.6: Average KL divergence between unimodal and multimodal latent distributions for the basic PoE model and the KL-regularized PoE model ($\lambda^{\star}_{\mathrm{KL}}$). For each model, the divergence is first averaged over the samples in each validation fold, and the resulting five fold-level values are summarized as mean ± standard deviation. Lower values indicate stronger alignment between the corresponding latent distributions.

Model	Total KL	joint → MRI	joint → FDG-PET
PoE basic	69.1308 ± 10.5201	41.8940 ± 7.3841	27.2368 ± 4.1721
PoE + KL	14.9953 ± 2.8997	8.3297 ± 1.7414	6.6656 ± 1.2256

To assess whether the KL term achieves its intended effect on the latent representations, Table 6.6 reports the average per-sample KL divergence between the multimodal and unimodal latent distributions, for both the basic and KL-regularized PoE models. For each model, this divergence is first averaged over the samples in each validation fold, and the resulting five fold-level values are then summarized as a mean and standard deviation. Total refers to the sum of both directions, joint → MRI and joint → FDG-PET. The total KL divergence decreases substantially, from 69.13 ± 10.52 in the basic model to 14.99 ± 2.90 with KL regularization, with a comparable reduction observed for both the MRI and FDG-PET components individually. This indicates that the regularizer is indeed successful at pulling the unimodal and multimodal latent distributions closer together.

This effect is also visible qualitatively. Figure 6.6 shows a shared t-SNE projection of the latent embeddings produced by the basic and KL-regularized PoE models, computed jointly so that relative distances between the two sets of embeddings remain meaningful. The KL-regularized embeddings for the MRI and FDG-PET pathways appear visibly closer to the multimodal region than their unregularized counterparts, supporting the interpretation that the KL term is effectively

reducing the separation between the modality-specific latent regions identified in Figure 6.5. Despite this alignment, however, the effect on classification performance is not uniform across pathways: while the MRI and multimodal pathways benefit from regularization, the FDG-PET pathway consistently worsens, suggesting that structural alignment in the latent space does not automatically translate into improved downstream performance for every modality.

Table 6.7: Test-set fold-wise comparison between the basic PoE model and the KL-regularized PoE model ($\lambda^{\star}_{\mathrm{KL}}$) on the MRI+FDG-PET pathway. For each fold, the reported values correspond to the model selected on that fold's validation split, evaluated once on the held-out test set. Δ denotes PoE+KL minus basic PoE; positive values indicate an improvement due to KL regularization.

Fold	Balanced accuracy			Macro F1-score		
	Basic	+KL	Δ	Basic	+KL	Δ
0	0.6262	0.6635	+0.0373	0.5216	0.5595	+0.0382
1	0.5860	0.6727	+0.0867	0.5119	0.5801	+0.0682
2	0.6643	0.6628	-0.0015	0.5432	0.5650	+0.0218
3	0.6607	0.6312	-0.0295	0.5748	0.5683	-0.0065
4	0.6763	0.6342	-0.0421	0.5868	0.5445	-0.0423
Mean ± std	0.6427 ± 0.0329	0.6529 ± 0.0169	+0.0108 ± 0.0524	0.5476 ± 0.0291	0.5635 ± 0.0117	+0.0159 ± 0.0423

A closer look at the fold-wise results for the multimodal pathway, reported in Table 6.7, further nuances the aggregate improvement described above. While the mean balanced accuracy and macro F1-score both increase under KL regularization, this improvement is not consistent across folds: balanced accuracy improves in only two of the five folds (0 and 1), and is essentially unchanged or worse in the remaining three, while macro F1-score improves in three folds and worsens in the other two. The overall positive average is therefore driven disproportionately by folds 0 and 1, rather than reflecting a uniform improvement across the cross-validation splits. This suggests that, while KL regularization provides a positive average effect on the multimodal pathway, this effect should be interpreted with some caution regarding its consistency.

6.4. Product-of-Experts with MetaReg

Table 6.8: Test-set balanced accuracy and macro F1-score of the Product-of-Experts model with MetaReg regularization ($\lambda_R = 1$), compared to the basic PoE model ($\lambda_R = 0$), across the three inference pathways. Results are reported as mean ± standard deviation across the five (one per fold) evaluations on the test set.

Model	MRI pathway	FDG-PET pathway	MRI + FDG-PET pathway
Balanced accuracy			
PoE basic	0.5118 ± 0.0286	0.6180 ± 0.0217	0.6427 ± 0.0329
PoE + MetaReg	**0.5250 ± 0.0263**	**0.6323 ± 0.0114**	**0.6600 ± 0.0122**
Macro F1-score			
PoE basic	0.4374 ± 0.0192	0.5319 ± 0.0180	0.5476 ± 0.0291
PoE + MetaReg	**0.4427 ± 0.0221**	**0.5501 ± 0.0113**	**0.5710 ± 0.0137**

Table 6.8 compares the test-set performance of the basic PoE model and the MetaReg-regularized model ($\lambda_R = 1$) across the three inference pathways. Unlike KL regularization, MetaReg improves performance over the basic PoE model consistently across all three pathways and both metrics: balanced accuracy improves by 1.3 points for MRI, 1.4 points for FDG-PET and 1.7 points for the multimodal pathway, with comparable gains in macro F1-score.

Comparing these results to the original single-pathway baselines (Table 6.2), the MRI pathway remains considerably behind the corresponding baseline, by 5.9 points of balanced accuracy and 9.0 points of macro F1-score, a gap similar in magnitude to that observed for the basic PoE model. The FDG-PET pathway, however, is now much closer to its baseline, with a gap of only 0.9 points of balanced accuracy and 1.6 points of macro F1-score. Most notably, the multimodal pathway

exceeds the MRI+FDG-PET baseline on both metrics, by 0.3 points of balanced accuracy and 2.2 points of macro F1-score, making this the only configuration considered so far in which a PoE variant outperforms the corresponding baseline on both metrics simultaneously.

Table 6.9: Fold-wise comparison between the basic PoE model and the MetaReg-regularized PoE model on the MRI+FDG-PET pathway. For each fold, the reported values correspond to the model selected on that fold's validation split, evaluated once on the held-out test set. Δ denotes PoE+MetaReg minus basic PoE; positive values indicate an improvement due to MetaReg regularization.

Fold	Balanced accuracy			Macro F1-score		
	Basic	+MetaReg	Δ	Basic	+MetaReg	Δ
0	0.6262	0.6442	+0.0180	0.5216	0.5491	+0.0275
1	0.5860	0.6526	+0.0666	0.5119	0.5701	+0.0582
2	0.6643	0.6628	-0.0015	0.5432	0.5881	+0.0449
3	0.6607	0.6806	+0.0199	0.5748	0.5823	+0.0075
4	0.6763	0.6598	-0.0165	0.5868	0.5654	-0.0214
Mean ± std	0.6427 ± 0.0329	0.6600 ± 0.0122	+0.0173 ± 0.0311	0.5476 ± 0.0291	0.5710 ± 0.0137	+0.0234 ± 0.0287

Given the relevance of this result, Table 6.9 reports the fold-wise comparison between the basic PoE model and the MetaReg-regularized model on the multimodal pathway. The improvement is considerably more consistent than the one observed for KL regularization in Table 6.7: balanced accuracy improves in three of the five folds and macro F1-score improves in four of the five folds, with only fold 4 showing a decline on both metrics. This suggests that, at least for the multimodal pathway, the benefit of MetaReg regularization is not driven by a small number of favourable folds to the same extent as observed for KL regularization.

6.5. Product-of-Experts with KL and MetaReg regularization

Table 6.10: Test-set balanced accuracy and macro F1-score of the Product-of-Experts model combining KL and MetaReg regularization ($\lambda^{\star}_{\mathrm{KL}}$, $\lambda_R = 1$), compared to the KL-only PoE model, across the three inference pathways. Results are reported as mean ± standard deviation across the five (one per fold) evaluations on the test set. The best value per column is shown in bold.

Model	MRI pathway	FDG-PET pathway	MRI + FDG-PET pathway
Balanced accuracy			
PoE + KL	**0.5553 ± 0.0238**	0.5923 ± 0.0393	0.6529 ± 0.0169
PoE + KL + MetaReg	0.5224 ± 0.0502	**0.6257 ± 0.0149**	**0.6621 ± 0.0136**
Macro F1-score			
PoE + KL	**0.4994 ± 0.0180**	0.5264 ± 0.0279	**0.5635 ± 0.0117**
PoE + KL + MetaReg	0.4456 ± 0.0386	**0.5330 ± 0.0188**	0.5560 ± 0.0253

Table 6.10 compares the test-set performance of the KL-regularized model and the model combining both KL and MetaReg regularization ($\lambda_{\mathrm{KL}} = \lambda^{\star}_{\mathrm{KL}} = 10^{-2}$, $\lambda_R = 1$). The effect of adding MetReg on top of KL regularization is mixed: the FDG-PET pathway improves, by 3.3 points of balanced accuracy and 0.7 points of macro F1-score, and the multimodal pathway improves slightly in balanced accuracy (+0.9 points) but declines slightly in macro F1-score (-0.75 points). The MRI pathway, however, worsens considerably, losing 3.3 points of balanced accuracy and 5.4 points of macro F1-score relative to KL regularization alone.

Table 6.11: Summary of test-set balanced accuracy and macro F1-score across all trained model variants and inference pathways. Results are reported as averages across the five (one per fold) evaluations on the test set. The best value per column and metric is shown in bold.

Model	Balanced accuracy			Macro F1-score		
	MRI	FDG-PET	MRI+FDG-PET	MRI	FDG-PET	MRI+FDG-PET
Baseline	**0.5835**	**0.6411**	0.6566	**0.5324**	**0.5656**	0.5490
PoE basic	0.5118	0.6180	0.6427	0.4374	0.5319	0.5476
PoE + KL	0.5553	0.5923	0.6529	0.4994	0.5264	0.5635
PoE + MetaReg	0.5250	0.6323	0.6600	0.4427	0.5501	**0.5710**
PoE + KL + MetaReg	0.5224	0.6257	**0.6621**	0.4456	0.5330	0.5560

A broader comparison across all trained variants, summarized in Table 6.11, confirms that the baseline models remain the best-performing option for both the MRI and FDG-PET pathways, on both balanced accuracy and macro F1-score: no combination of regularizers is sufficient to close this gap entirely. Among the PoE variants themselves, however, KL regularization alone achieves the best MRI performance, while MetaReg alone achieves the best FDG-PET performance, on both metrics in each case. It is only in the multimodal pathway that PoE variants outperform the baseline: the combined KL+MetaReg model achieves the highest balanced accuracy overall (0.6621), while MetaReg alone achieves the highest macro F1-score overall (0.5710). Notably, combining KL and MetaReg does not achieve the best result for either metric in the unimodal pathways, and only marginally improves on MetaReg alone in the multimodal balanced accuracy.

These results suggest that KL regularization and MetaReg do not simply combine additively. A possible reason is that the MetaReg regularizer R_ψ is meta-learned in Algorithm 1 without any KL regularization present. When the two are then combined in the final training run, the classifier regularizer is therefore applied under a different condition than the one it was meta-learned under, namely one where the latent distributions are also being reshaped by the KL term. This mismatch between the condition used to learn R_ψ and the condition used to apply it may explain the lack of a consistent additional benefit. This explanation is speculative, and we leave a more detailed investigation of this interaction, for instance by meta-learning R_ψ jointly with KL regularization active, to future work.

Overall, among the regularization strategies considered in this thesis, MetaReg alone provides the most consistent improvement over the basic PoE model, achieving the best or near-best macro F1-score across all three pathways and being the only variant that outperforms the corresponding baseline on both metrics in the multimodal setting, as discussed in Section 6.5. KL regularization alone remains preferable specifically for the MRI pathway. Combining the two regularizers does not provide a further consistent benefit over using either regularizer individually.

7

Conclusion

This thesis addressed the problem of diagnosing Alzheimer's disease from neuroimaging data, focusing on the three-class classification task that separates cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects. This task is a standard but notoriously difficult benchmark on ADNI data, mainly because the MCI class is clinically heterogeneous and because the number of subjects is small relative to the dimensionality of the images. Most multimodal methods in the literature, including recent state-of-the-art models, assume that both MRI and FDG-PET are available for every subject at training and inference time. This assumption rarely holds in practice, since PET is far less common than MRI both in clinical settings and within ADNI itself. The central goal of this work was therefore to design a model that can use all available data and remain usable while also studying how such a model should be trained.

To this end, three main contributions were developed. First, a Product-of-Experts model was formulated, in which a shared latent variable represents the disease-relevant brain state and each modality contributes a Gaussian expert to a common latent distribution. Through an MVAE-inspired reparametrization, the model produces closed-form unimodal and multimodal latent distributions, and can generate a prediction from MRI alone, FDG-PET alone, or both modalities combined, always using the same shared classifier. In this way, the missing-modality problem is handled by design, without imputing the absent modality. Second, two regularization strategies were introduced to encourage the shared classifier to behave reliably across the different latent-input regimes induced by different modality-availability patterns: a meta-learned classifier regularizer (MetaReg), inspired by domain-generalization methods, and a Kullback-Leibler (KL) alignment term that pulls the unimodal and multimodal latent distributions closer together. Third, the resulting framework was compared against unimodal and multimodal baselines, and its components were studied through a set of ablation experiments.

7.1. Main findings

A first outcome concerns the way the predictive distribution is approximated. The deterministic mean approximation consistently outperformed the stochastic Monte Carlo approximation in terms of validation balanced accuracy across all three baselines. The stochastic objective tended to collapse toward predicting the MCI class, especially in the FDG-PET and multimodal settings, which are trained on the smallest sets. This behaviour appeared milder for the MRI baseline, which is trained on the largest amount of data. While this is only a hypothesis, since the three settings differ in more than just sample size, it suggests that the stochastic approximation is more sensitive to limited data. For this reason, the deterministic approximation was adopted throughout the thesis.

A second, and more central, finding is that jointly training the Product-of-Experts model with a

plain supervised objective does not, by itself, improve performance. In its basic form, the model was outperformed by the dedicated unimodal and multimodal baselines on every inference pathway. The gap was largest for the MRI pathway and smallest for the multimodal pathway. Part of this asymmetry may be explained by the mini-batch construction: since one epoch is defined as a full pass over the smallest set, and the MRI-available set is much larger, each individual MRI sample is effectively seen fewer times than each FDG-PET or complete sample. A t-SNE analysis of the latent means showed that the MRI-only, FDG-PET-only and multimodal encodings form three clearly separated clusters, with no overlap. This confirms the concern that motivated the regularization strategies: because the same classifier is applied to all three regimes, a decision boundary that suits one region of the latent space may be poorly suited to another.

The two regularizers addressed this concern with different degrees of success. KL regularization strongly reduced the divergence between the unimodal and multimodal latent distributions, confirming that it aligns the representations as intended. This alignment improved the MRI and multimodal pathways, but slightly harmed the FDG-PET pathway, showing that reducing the geometric separation of the latent regions does not automatically improve classification for every modality. Moreover, the improvement on the multimodal pathway was not uniform across folds, and was driven mainly by two of the five folds. MetaReg, in contrast, improved performance over the basic model consistently across all three pathways and both metrics, and did so more evenly across folds. Most notably, the MetaReg model was the only single-regularizer configuration whose multimodal pathway exceeded the corresponding multimodal baseline on both balanced accuracy and macro F1-score.

Combining the two regularizers did not yield an additional consistent benefit. The combined model achieved the highest multimodal balanced accuracy overall, but did not improve on MetaReg alone in macro F1-score, and clearly worsened the MRI pathway. A plausible reason is that the MetaReg regularizer is meta-learned in a setting where the KL term is absent, and is then applied during final training under a different condition, in which the latent distributions are simultaneously reshaped by the KL term. This mismatch between the condition used to learn the regularizer and the condition used to apply it may explain why the two techniques do not simply add up. Taken together, the ablations indicate that MetaReg provides the most consistent improvement over the basic Product-of-Experts model, that KL regularization is preferable specifically for the MRI pathway, and that the dedicated baselines remain the strongest option for the two unimodal pathways.

7.2. Comparison with the state of the art

It is useful to place these results next to DiaMond [1], the recent transformer-based model that fuses MRI and FDG-PET and reports a balanced accuracy of 65.2% and a macro F1-score of 64.9% on the same three-class task. On balanced accuracy, the multimodal pathway of the proposed framework is competitive with this reference: the multimodal baseline reaches 65.66%, and the best regularized variant (KL combined with MetaReg) reaches 66.21%, with MetaReg alone close behind at 66.00%. On macro F1-score, however, all the models studied here remain clearly below DiaMond, with the best value being 57.10% for MetaReg. This pattern, in which balanced accuracy is comparable but macro F1-score is not, suggests that the present models recover the recall of each class reasonably well but suffer more from over-prediction, which the precision term of the F1-score penalizes.

This comparison is contextual rather than a controlled head-to-head experiment, since DiaMond was evaluated under its own preprocessing, temporal-matching and splitting choices, which can change the resulting dataset and the reported numbers. Even so, the result is encouraging. As reported in Section 5.2, the deployed Product-of-Experts model contains 334,219 trainable parameters, against the 30,106,015 of DiaMond, so a comparable level of balanced accuracy is reached with a model roughly 90 times smaller. Unlike DiaMond, moreover, the proposed model remains usable when only one modality is available at inference time. In this sense, the framework trades some macro F1-score performance for robustness to missing modalities and for a considerably smaller model, which is the trade-off it was built to make.

7.3. Limitations

Several limitations should be kept in mind when interpreting these results. The most important is the small size of the dataset, in particular the very small number of unpaired FDG-PET samples. This scarcity makes the FDG-PET and multimodal pathways prone to unstable training and limits the strength of any conclusion drawn from them. A second limitation concerns the temporal matching used to build the dataset. A window of 180 days was needed to obtain enough samples, but such a large window may violate the assumption that the diagnosis and the neuroimaging scans describe the same clinical state, especially for a progressive disease. Related to this, both 1.5T and 3T MRI scans were treated as a single modality, which increases the amount of usable data but also introduces heterogeneity unrelated to the disease.

Further limitations follow from the modelling assumptions. The conditional independence assumption underlying the Product-of-Experts fusion rule is idealized and is likely violated in practice by scanner effects, site-specific noise and preprocessing artefacts shared by the two modalities. The Gaussian parametrization with diagonal covariance keeps the model tractable, but may not capture the true structure of the uncertainty in the latent space. On the training side, the mini-batch construction disadvantages the MRI pathway, as discussed above, and the interaction between the two regularizers was not fully controlled, since the MetaReg regularizer was meta-learned without the KL term active. Finally, several improvements over the basic model were not uniform across cross-validation folds, so the average gains should be interpreted together with their variability rather than in isolation.

7.4. Future work

The most direct follow-up is to study the interaction between the two regularizers more carefully, for instance by meta-learning the MetaReg regularizer while the KL term is already active, so that it is learned under the same condition in which it is later applied. On the data side, a stricter temporal-matching criterion, or an explicit modelling of the time gap between scans and diagnosis, could reduce the label noise introduced by the current window. Modelling scanner and site effects, or relaxing the conditional independence assumption, may make the latent representation more faithful to the true data-generating process. Richer latent distributions, such as full covariance matrices or non-Gaussian families, could also be explored, at the cost of losing the closed-form expressions used here.

A different line of work concerns the training procedure itself. A more balanced mini-batch construction, which exposes the MRI pathway to the full diversity of its data as often as the other pathways, might recover part of the performance lost by the MRI expert under joint training. Given that the models here reach a balanced accuracy comparable to the state of the art but a lower macro F1-score, future work could also focus specifically on the precision of the predictions, for example through class-balancing or calibration techniques. Finally, the Product-of-Experts formulation extends naturally to more than two modalities, so incorporating additional sources of information available in ADNI, such as amyloid or tau PET, cerebrospinal fluid biomarkers or cognitive scores, is a promising way to improve diagnostic performance while preserving the ability to handle missing inputs.

Bibliography

[1] Yitong Li, Morteza Ghahremani, Youssef Wally, and Christian Wachinger. Diamond: Dementia diagnosis with multi-modal vision transformers using mri and pet. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 107–116. IEEE, 2025.

[2] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[4] Richard J Caselli, Thomas G Beach, David S Knopman, and Neill R Graff-Radford. Alzheimer disease: scientific breakthroughs and translational challenges. In *Mayo Clinic Proceedings*, volume 92, pages 978–994. Elsevier, 2017.

[5] Jorge A Trejo-Lopez, Anthony T Yachnis, and Stefan Prokop. Neuropathology of alzheimer's disease. *Neurotherapeutics*, 19(1):173–185, 2022.

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

[7] Premkumar Shantilal Baviskar and Hitendra Shaligram Mahajan. Unveiling alzheimer's disease (1901–2025): Historical insights, global burden, biological mechanisms, diagnostics, and therapeutic strategies. *Ageing Research Reviews*, 114:102990, 2026.

[8] Kaj Blennow and Henrik Zetterberg. Biomarkers for alzheimer's disease: current status and prospects for the future. *Journal of internal medicine*, 284(6):643–663, 2018.

[9] Robert Katzman. The prevalence and malignancy of alzheimer disease: a major killer. *Archives of neurology*, 33(4):217–218, 1976.

[10] Rishi S Madnani. Alzheimer's disease: a mini-review for the clinician. *Frontiers in Neurology*, 14:1178588, 2023.

[11] Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323, 2016.

[12] WHO. Dementia, 2025.

[13] Lyn Xuan Tay, Siew Chin Ong, Lynn Jia Tay, Trecia Ng, and Thaigarajan Parumasivam. Economic burden of alzheimer's disease: a systematic review. *Value in health regional issues*, 40:1–12, 2024.

[14] Scott Ayton and Ashley I. Bush. beta-amyloid: The known unknowns. *Ageing Research Reviews*, 65:101212, 2021.

[15] John A Hardy and Gerald A Higgins. Alzheimer's disease: the amyloid cascade hypothesis. *Science*, 256(5054):184–185, 1992.

[16] Christian Behl. In 2024, the amyloid-cascade-hypothesis still remains a working hypothesis, no less but certainly no more. *Frontiers in Aging Neuroscience*, 16:1459224, 2024.

[17] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[18] Glenda Halliday. Pathology and hippocampal atrophy in alzheimer's disease. *The Lancet Neurology*, 16(11):862–864, 2017.

[19] A. Berger. How does it work?: Magnetic resonance imaging. *BMJ*, 2002.

[20] Martin John Graves. 3 t: the good, the bad and the ugly. *The British Journal of Radiology*, 95(1130):20210708, 2022.

[21] Kenji Kawada, Masayoshi Iwamoto, and Yoshiharu Sakai. Mechanisms underlying 18f-fluorodeoxyglucose accumulation in colorectal cancer. *World journal of radiology*, 8(11):880, 2016.

[22] S Shokouhi, D Claassen, and WR Riddle. Imaging brain metabolism and pathology in alzheimer's disease with positron emission tomography. *Journal of Alzheimer's disease & Parkinsonism*, 4(2):143, 2014.

[23] Marla Narazani, Ignacio Sarasua, Sebastian Pölsterl, Aldana Lizarraga, Igor Yakushev, and Christian Wachinger. Is a pet all you need? a multi-modal study for alzheimer's disease using 3d cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2022.

[24] Jorge Samper-González, Ninon Burgos, Sabrina Fontanella, Hugo Bertin, Marie-Odile Habert, Stanley Durrleman, Theodoros Evgeniou, Olivier Colliot, and Alzheimer's Disease Neuroimaging Initiative. Yet another adni machine learning paper? paving the way towards fully-reproducible research on classification of alzheimer's disease. In *International Workshop on Machine Learning in Medical Imaging*, pages 53–60. Springer, 2017.

[25] Ross A Dunne, Dag Aarsland, John T O'Brien, Clive Ballard, Sube Banerjee, Nick C Fox, Jeremy D Isaacs, Benjamin R Underwood, Richard J Perry, Dennis Chan, et al. Mild cognitive impairment: the manchester consensus. *Age and ageing*, 50(1):72–80, 2021.

[26] Jorge Samper-González, Ninon Burgos, Simona Bottani, Sabrina Fontanella, Pascal Lu, Arnaud Marcoux, Alexandre Routier, Jérémy Guillon, Michael Bacci, Junhao Wen, et al. Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to mri and pet data. *NeuroImage*, 183:504–521, 2018.

[27] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[28] Rong Zhang, Jinhua Sheng, Qiao Zhang, Junmei Wang, and Binbing Wang. A review of multimodal fusion–based deep learning for alzheimer's disease. *Neuroscience*, 576:80–95, 2025.

[29] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2302–2310, 2021.

[30] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[31] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.

[32] Mike Wu and Noah D. Goodman. Multimodal generative models for scalable weakly-supervised learning. *CoRR*, abs/1802.05335, 2018.

[33] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[34] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[35] Tianyi Liu, Zhaorui Tan, Haochuan Jiang, Xi Yang, and Kaizhu Huang. Mind the gap: Promoting missing modality brain tumor segmentation with alignment. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3531–3536. IEEE, 2024.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[37] Yuta Oshima, Masahiro Suzuki, and Yutaka Matsuo. Enhancing unimodal latent representations in multimodal vaes through iterative amortized inference. *New Generation Computing*, 44(2):17, 2026.

[38] Tim M. Tierney, Nicholas A. Alexander, John Ashburner, Nicole Labra Avila, Yaël Balbastre, Gareth Barnes, Yulia Bezsudnova, Mikael Brudfors, Korbinian Eckstein, Guillaume Flandin, Karl Friston, Amirhossein Jafarian, Olivia S. Kowalczyk, Vladimir Litvak, Johan Medrano, Stephanie Mellor, George O'Neill, Thomas Parr, Adeel Razi, Ryan Timms, and Peter Zeidman. Spm 25: open source neuroimaging analysis software. *Journal of Open Source Software*, 10(110):8103, 2025.

[39] John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

[40] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[44] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.