UNIVERSITY OF PADUA

DEPARTMENT OF INFORMATION ENGINEERING

Master's Degree in Bioengineering

"Interpretable Deep Learning for MRI-Based Content-Based Image Retrieval in Alzheimer's Disease"

Supervisor:

Prof. Mattia Veronese

Candidate:

Camilla Sabbion

Student ID: 2157134

Co-supervisors:

Prof. Rodrigo Moreno Serrano

Dr. Félix Nieto del Amor

Dr. Sanna Persson

ACADEMIC YEAR 2025/2026

Abstract

Alzheimer's disease is a neurodegenerative disorder in which structural brain alterations, observable through magnetic resonance imaging (MRI), play a key role in both diagnosis and disease monitoring. In recent years, deep learning approaches have shown great potential for automatically extracting meaningful patterns from MRI data. However, many of these models rely on heavily preprocessed images and behave as black boxes, limiting their clinical applicability and interpretability.

In this context, this thesis develops and evaluates a deep learning framework capable of reproducing NeuroCBIR target embeddings directly from minimally preprocessed T1-weighted MRI scans. A conventional 3D CNN was used as a baseline and compared with a prototype-based PIP-Net model adapted for embedding regression. Both models were evaluated through subject-level image retrieval on OASIS-3 and on three external datasets.

Across the four validation folds, PIP-Net achieved substantially higher and more stable retrieval performance than the CNN baseline, with an mS@1 of 0.865 ± 0.007, compared with 0.613 ± 0.157 for the CNN.

On the external datasets, PIP-Net achieved mS@1 values of 0.994 on MIRIAD, 0.827 on AIBL, and 0.604 on SLIM, compared with 0.995, 0.696, and 0.381 for the CNN baseline, indicating strong generalization to datasets with characteristics similar to the training data, with reduced performance under larger domain shifts. Overall, the findings suggest that NeuroCBIR-like embeddings can be approximated from minimally preprocessed MRI scans, while prototype activations provide an additional way to inspect the image regions contributing to the predicted representation.

Keywords

Alzheimer's disease; Content-Based Image Retrieval; Magnetic Resonance Imaging; Deep Learning; Explainable Artificial Intelligence; Prototype-Based Networks.

Riassunto

La malattia di Alzheimer è una patologia neurodegenerativa nella quale le alterazioni strutturali del cervello, osservabili mediante risonanza magnetica (MRI), rivestono un ruolo centrale sia nella diagnosi sia nel monitoraggio della progressione della malattia. Negli ultimi anni, gli approcci di deep learning hanno mostrato un notevole potenziale nell'estrazione automatica di pattern significativi dai dati di risonanza magnetica. Tuttavia, molti di questi modelli richiedono immagini sottoposte a procedure di preprocessing complesse e si comportano come sistemi black-box, limitandone l'applicabilità clinica e l'interpretabilità.

In questo contesto, la presente tesi sviluppa e valuta un framework di deep learning in grado di riprodurre gli embedding target di NeuroCBIR direttamente a partire da immagini MRI T1-pesate sottoposte a un preprocessing minimo. Come modello di riferimento è stata utilizzata una CNN 3D convenzionale, confrontata con PIP-Net, un modello basato su prototipi e adattato a un compito di regressione degli embedding. Entrambi i modelli sono stati valutati mediante image retrieval a livello di soggetto sul dataset OASIS-3 e su tre dataset esterni.

Considerando i quattro fold di validazione, PIP-Net ha ottenuto prestazioni di retrieval nettamente superiori e più stabili rispetto alla CNN di baseline, raggiungendo un valore di mS@1 pari a 0.865 ± 0.007, rispetto a 0.613 ± 0.157 ottenuto dalla CNN. Sui dataset esterni, PIP-Net ha raggiunto valori di mS@1 pari a 0.994 su MIRIAD, 0.827 su AIBL e 0.604 su SLIM, rispetto rispettivamente a 0.995, 0.696 e 0.381 ottenuti dalla CNN di baseline. Questi risultati indicano una buona capacità di generalizzazione su dataset con caratteristiche simili a quelle dei dati di training, mentre le prestazioni si riducono in presenza di differenze di dominio più marcate. Nel complesso, i risultati suggeriscono che sia possibile approssimare embedding simili a quelli di NeuroCBIR a partire da immagini MRI minimamente preprocessate. Le attivazioni dei prototipi forniscono inoltre uno strumento aggiuntivo per analizzare le regioni dell'immagine che contribuiscono alla rappresentazione predetta.

Parole chiave

Malattia di Alzheimer; content-based image retrieval; risonanza magnetica; deep learning; intelligenza artificiale; reti neurali.

CONTENTS

List of Figures

LIST OF TABLES

1. Introduction

Alzheimer's disease (AD) represents one of the most urgent public health challenges associated with population ageing. Approximately 57 million people worldwide in 2021 are affected by dementia, with nearly 10 million new cases every year, and Alzheimer's disease is estimated to account for 60–70% of all dementia cases [1]. Dementia is not only one of the major causes of disability and dependency among older adults, but it is also the seventh leading cause of death globally and the burden of the disease is expected to increase substantially in the coming decades. The number of people living with dementia is projected to reach 78 million by 2030 and 139 million by 2050, largely reflecting the increasing number of older adults worldwide [2].

Despite its growing impact, the early recognition of Alzheimer's disease remains clinically challenging. Initial symptoms, such as memory difficulties, reduced judgement, or subtle changes in reasoning, may be mild and progressive, and can therefore be difficult to distinguish from normal age-related cognitive changes. Mild cognitive impairment (MCI) further complicates this distinction, as individuals may show memory or thinking problems greater than expected for their age, while still maintaining independence in daily activities. As a result, Alzheimer's disease is often diagnosed only after cognitive decline has become more evident, limiting the opportunity for timely clinical management [3].

In this context, neuroimaging plays a central role in supporting the clinical assessment of neurodegenerative diseases. Structural magnetic resonance imaging can reveal anatomical patterns associated with Alzheimer's disease, including progressive brain atrophy in regions involved in memory and cognition. However, the increasing availability of large MRI databases makes the manual identification of comparable cases inefficient and difficult. Content-Based Image Retrieval (CBIR) systems address this problem by retrieving images that are similar to a given query based on intrinsic image features rather than metadata. In neuroimaging, this approach can support clinicians and researchers by providing rapid access to structurally comparable cases, facilitating clinical decision-making, differential diagnosis, and population-level analysis.

Therefore, the goal of this thesis is to explore an interpretable deep learning approach for MRI-based CBIR in Alzheimer's disease. In this way, the system is intended as a clinical support

tool to make the evaluation of MRI scans faster, more transparent, and easier to compare with previous cases.

Although deep learning has shown strong potential for MRI-based analysis, many neuroimaging pipelines still depend on extensive preprocessing before images can be used for quantitative analysis. Structural MRI preprocessing commonly includes operations such as bias field correction, skull stripping, spatial registration, intensity normalization, and anatomical segmentation. These steps are important because they reduce acquisition-related variability and help ensure that subsequent models focus on brain anatomy rather than irrelevant image artefacts. For example, FreeSurfer provides a complete MRI processing stream including skull stripping, bias field correction, registration, anatomical segmentation, cortical reconstruction, and parcellation [4].

However, this dependence on preprocessing also represents a practical limitation. Structural MRI preprocessing can introduce a substantial computational overhead, particularly when applied to large-scale datasets. Although a complete FreeSurfer processing pipeline may require several hours per subject, NeuroCBIR does not require the full pipeline; the preprocessing steps needed to obtain its input images take approximately 30 minutes per scan. This additional processing time can still become a relevant bottleneck when thousands of MRI scans need to be analyzed. Therefore, while preprocessing improves anatomical standardization, it may also reduce the scalability and immediate usability of AI-based tools in clinical or research workflows [5].

This issue is particularly relevant for content-based image retrieval systems in neuroimaging. CBIR aims to retrieve images similar to a given query based on intrinsic image features rather than metadata, allowing clinicians and researchers to rapidly access comparable cases. In this context, NeuroCBIR represents a recent example of this approach, providing fast whole-brain and region-specific retrieval of 3D T1-weighted MRI scans [6]. Its embeddings are designed to capture anatomical similarity and are evaluated using subject re-identification and other downstream tasks. Nevertheless, NeuroCBIR requires MRI preprocessing, including bias field correction, skull stripping, and spatial registration, which motivates the question of whether similar retrieval behaviour can be approximated directly from minimally preprocessed scans. To address this gap, this thesis investigates whether NeuroCBIR-like embeddings can be predicted from minimally preprocessed T1-weighted MRI images. The main objective is to develop and evaluate deep learning models that approximate the NeuroCBIR embedding space while reducing the dependence on time-consuming preprocessing. The project therefore does not aim to replace clinical diagnosis, but to explore a support tool that may help clinicians retrieve similar MRI cases more efficiently and transparently.

2. Background

2.1 Alzheimer's Disease

Alzheimer's disease (AD) is the primary cause of dementia worldwide and represents a major contributor to its global clinical and societal burden [7]. Distinguishing AD from normal aging is challenging especially in the early stages, since both are characterized by cognitive decline and neuropathological changes [8].

The earliest symptoms of AD typically involve deficits in anterograde episodic and prospective memory while later stages affect working memory, spatial orientation, and awareness of one's actions. From a neuropathological perspective, AD is characterized by the accumulation of amyloid plaques and neurofibrillary tangles, which contribute to progressive neurodegeneration and macroscopic brain atrophy. This degeneration primarily affects medial temporal and parietal regions, as well as the cingulate cortex and frontal lobes, while primary sensory, motor, and visual areas are generally less affected during the course of the disease [9].

Accordingly, it is crucial to understand the neuropathological mechanisms underlying Alzheimer's disease in order to support its clinical identification, as well as the identification of Mild Cognitive Impairment.

The principal advancements in clinical research on Alzheimer's disease focus on biomarkers capable of detecting amyloid plaques and neurofibrillary tau tangles directly in the body. In this direction, several techniques have been developed, including the analysis of cerebrospinal fluid and positron emission tomography (PET) imaging. A crucial finding that has emerged from Alzheimer's disease research is the existence of a long preclinical stage, lasting up to two decades, during which amyloid plaques and tau tangles accumulate in the brain without patients showing evident cognitive decline.

However, these techniques can be invasive, expensive, or not always easily accessible in clinical practice. For this reason, non-invasive neuroimaging approaches have gained increasing importance.

2.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging is a non-invasive medical imaging technique widely used in radiology to generate detailed images of the anatomy and physiological processes of the human body and has played a fundamental role in this context.

MRI images are obtained by measuring the relaxation properties of hydrogen protons after excitation by a radiofrequency (RF) pulse at the Larmor frequency. Two main types of relaxation are considered; longitudinal relaxation (T1) and transverse relaxation (T2), which reflect different tissue characteristics [10-11].

T1-weighted images primarily highlight anatomical structures and are especially sensitive to tissues with high fat content, due to that, fat-rich tissues, such as subcutaneous fat and bone marrow, appear bright in T1-weighted images. On the other hand, cerebrospinal fluid (CSF), which contains relatively little fat, appears dark. This contrast makes T1-weighted MRI particularly suitable for analyzing brain anatomy and identifying structural changes associated with neurodegenerative diseases [10-11].

Structural MRI techniques reveal characteristic patterns of brain alterations that enable clinicians to distinguish Alzheimer's disease from other neurological conditions. Furthermore, MRI can identify structural brain abnormalities associated with the risk of progression from mild cognitive impairment to Alzheimer's disease and with other behavioral and cognitive outcomes [12].

The Braak staging is a model used to describe the progression of brain atrophy in Alzheimer's disease, characterizing the spatial and temporal spread of neurodegeneration across the brain. For instance, in the early stages, structural alterations are primarily observed in two regions crucial for memory formation and cognitive processing, the medial temporal lobe (MTL), and the hippocampus. Moreover, neuroimaging studies have shown that, compared to healthy individuals, patients with Alzheimer's disease present a significant reduction in hippocampal volume and an even greater reduction in the entorhinal cortex. Individuals affected by Mild Cognitive Impairment typically show intermediate levels of medial temporal lobe atrophy, representing a transitional stage between normal aging and Alzheimer's disease. Moreover, diffuse hippocampal atrophy has been associated with impairments in memory and executive functioning in AD patients [10].

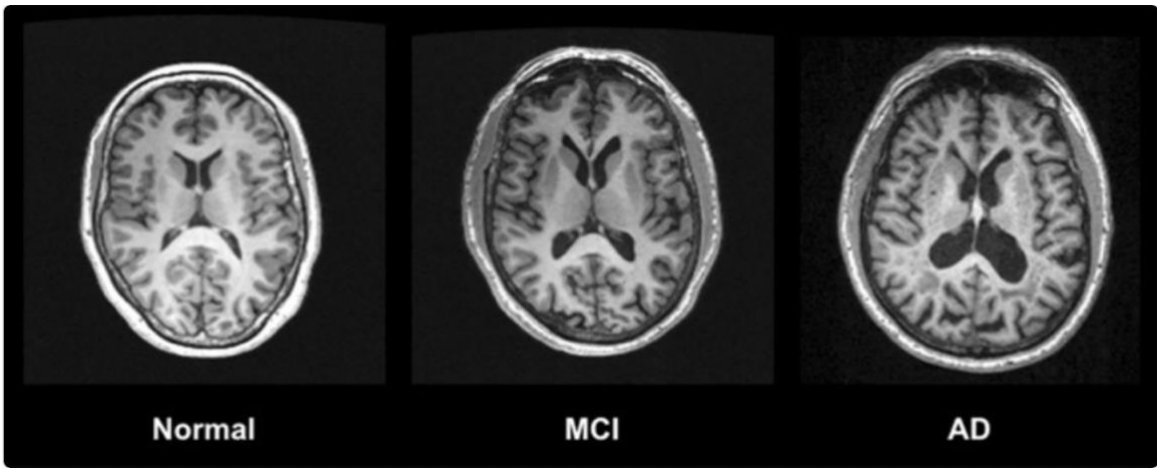

Figure 1: *T1-weighted MRI imaging using an MPRAGE (Magnetisation Prepared Rapid Gradient Echo) sequence shows decreased GM volume in an AD patient compared to a healthy control and intermediate GM decline in a patient with Mild cognitive impairment [10].*

2.3 Deep Learning

Deep learning is a branch of machine learning based on artificial neural networks composed of multiple layers. In comparison with traditional machine learning approaches, where relevant image descriptors are often manually defined, deep learning models can learn informative features directly from the input data. This property is particularly useful in medical image analysis, where anatomical and pathological patterns may be complex, and difficult to describe using a limited set of predefined measurements.

2.3.1 Convolutional Neural Network

Convolutional neural networks (CNNs) are among the most widely used deep learning architectures for image analysis. A CNN works with a sequence of convolutional layers, activation functions, pooling operations, and, depending on the task, fully connected layers. The main idea is that small learnable filters are applied to local regions of the image and detect simple local patterns in the first layers, such as edges, intensity changes, or textures, while deeper layers combine these local features into more abstract and task-specific representations [13-14-15-16].

The training of CNNs generally relies on backpropagation and gradient descent. Backpropagation makes it possible to compute how much each weight in the network contributes to the loss, by propagating the error backwards from the final layer and applying the chain rule. Once these gradients have been obtained, gradient descent updates the model

parameters in the direction that reduces the objective function. Through repeated iterations of this process, the network progressively learns a set of weights that improves its predictions on the training data [14–16].

In neuroimaging, CNNs are particularly relevant because they can be extended to three-dimensional data and applied directly to volumetric MRI scans. This makes them suitable for learning spatial patterns associated with neurodegeneration. For example, a previous study [17] proposed an interpretable framework based on an ensemble of 3D convolutional neural networks. By introducing the P-score as a digital biomarker, the model was able to identify interconnected neurodegenerative trajectories and highlight the early involvement of specific brain regions. These results were consistent with established models of Alzheimer's disease progression, such as Braak staging, and showed the potential of deep learning for non-invasive and personalized in-vivo staging of the disease.

2.3.2 Residual Networks

As convolutional neural networks grew deeper, image recognition results kept improving, but not indefinitely. Beyond a certain depth, plain networks become harder to optimize, and training error can actually rise rather than fall.

Residual Network (ResNet) was designed to introduce residual blocks with identity skip connections.

Rather than forcing a stack of layers to learn a target mapping H(x) directly, the layers are made to learn only the residual F(x), the gap between that target and the input itself. A residual block's output is then

$$H(x) = F(x) + x \tag{1}$$

where x is the block's input and F(x) the mapping learned by the convolutional layers; the shortcut connection simply adds x back to that output, letting the original signal skip over one or more layers unchanged (Figure 2) [18, 19].

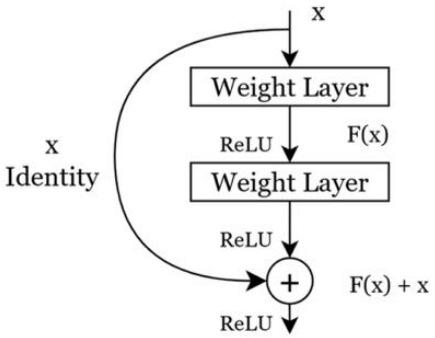

Figure 2: *Residual Block example.*

The key idea of residual learning is that a block does not need to learn the full transformation from the input to the output. Instead, it learns only a residual correction. When the optimal transformation is close to the identity, the network can simply make F(x) close to zero and preserve the input through the shortcut. In this way, deeper networks are easier to optimize and help maintain information flow across layers [18-19].

This idea is relevant in medical imaging because deep convolutional backbones are often used to extract image features. In the case of brain MRI, the same principle can be applied with three-dimensional convolutions, allowing the model to process the scan as a volumetric image while keeping the network trainable.

2.3.3 Deep Learning for Regression

Regression is a supervised learning task in which a model is trained to estimate continuous target values from input data instead of assigning a class. In medical imaging, regression has been widely used for tasks such as brain age estimation, disease severity prediction, and the estimation of quantitative biomarkers from MRI data [20].

Deep learning models, and in particular convolutional neural networks, can learn image representations directly from volumetric inputs and map them to numerical outputs. In a standard deep regression framework, an input image x is processed by a neural network f_θ, which produces a prediction $\hat{y} = f_\theta(x)$. The network parameters θ are optimized by minimizing the discrepancy between the predicted output and the reference target, commonly using losses such as mean absolute error or mean squared error.

In this thesis, the regression target is not a single scalar variable, such as age, but a continuous embedding vector extracted from the NeuroCBIR framework [6]. The goal is not only to reduce the numerical error between predicted and target embeddings, but also to preserve the similarity

structure of the original embedding space, since this structure determines the quality of the subsequent image retrieval step.

2.4 Content-Based Image Retrieval in Neuroimaging

Content-based image retrieval (CBIR) aims to retrieve images that are visually or anatomically similar to a query image by comparing learned feature representations rather than relying on manually defined descriptors. Early deep learning approaches mainly used convolutional neural networks as feature extractors, obtaining image embeddings from the final fully connected layers and comparing them using similarity measures [21].

2.4.1 NeuroCBIR Framework

Building upon these developments, the present thesis takes inspiration from the system NeuroCBIR [6]. NeuroCBIR is a CBIR framework designed to retrieve the most anatomically similar MRI scans given an incoming MRI query. The system relies on preprocessed T1-weighted MRI scans, including bias field correction, skull-stripping, and registration to the MNI305 space, using FreeSurfer v5.3.0 and the preprocessed TheHiveDB platform [22-23]. This preprocessing was intended to ensure that the embeddings captured brain-specific information and reflected anatomical similarity.

The first training stage of the NeuroCBIR framework consists of extracting latent representations from the input images using a variational autoencoder (VAE). These latent representations are then passed to an encoder-projector model trained with contrastive learning, with the aim of bringing scans from the same subject closer in latent space while pushing scans from different subjects farther apart. The resulting compact embeddings are then used to compare a query scan with the reference database and retrieve the most similar cases.

2.4.2 Practical Limitations of NeuroCBIR

Although NeuroCBIR shows the potential of CBIR for retrieving anatomically similar MRI scans, its use still depends on a specific preprocessing pipeline. This is relevant because, in a CBIR system, both the query scan and the reference database must be processed in a consistent way. Therefore, changes in preprocessing tools, parameters, or software versions may affect scalability and reproducibility. The NeuroCBIR study also notes that preprocessing steps, such

as skull-stripping, help prevent the embeddings from capturing non-brain information, but add time and complexity to the workflow [6].

Another limitation concerns interpretability. NeuroCBIR includes a preliminary analysis of the learned representations, but the retrieval process is still based on embeddings produced by deep neural networks. For this reason, the connection between the input MRI scan, the learned representation, and the retrieved cases is not fully explicit.

This motivates the direction followed in the present thesis: approximating NeuroCBIR-like embeddings from less extensively preprocessed MRI scans using an interpretable deep learning model. This choice is also supported by previous work such as DeepBrainNet, where Bashyam et al. showed that robust MRI-based biomarkers can be learned from T1-weighted scans using only minimal automated preprocessing, avoiding more demanding steps such as detailed segmentation, deformable registration, and advanced harmonization [24].

2.5 Explainable Artificial Intelligence

A major limitation of deep learning in medical applications is the difficulty of understanding how a model reaches its output. Although deep neural networks can achieve high predictive performance, their internal representations are often complex and not directly interpretable. In healthcare, this lack of transparency is particularly problematic, because a prediction should ideally be supported by evidence that clinicians can inspect and relate to meaningful clinical or anatomical patterns [25-27].
Explainable artificial intelligence (XAI) has emerged to address this problem by making machine learning models and their predictions more understandable.

2.5.1 Interpretability and XAI Methods

Several distinctions are commonly used in the literature when talking about explainable AI. The XAI methods can be classified into four categories based on: i) scope of explanations, ii) stages of implementation, iii) applicability to models, and iv) forms of explanation [27].
Regarding scope, explanations can be local, when they clarify a single prediction, or global, when they describe the overall behaviour of the model. Regarding timing, interpretability can be intrinsic, or ante-hoc, when it is built into the model architecture, or post-hoc, when an

explanation is produced after training a black-box model. Regarding applicability, some methods are model-specific, because they rely on internal elements of a given architecture, such as gradients or feature maps, whereas model-agnostic methods explain predictions without depending on the model structure. Finally, explanations can be presented in different formats, including numerical scores, visual heatmaps, textual descriptions, or rule-based statements[27].

In a medical context, interpretability has an additional relevance. A model prediction is more useful when it can be connected to image patterns that are compatible with known disease-related changes. For instance Viswan et al. [26] review XAI approaches used in Alzheimer's disease detection and report the use of several post-hoc techniques, including LIME, SHAP, Grad-CAM, and Layer-wise Relevance Propagation [26-27]. These methods can help inspect trained models and highlight regions or features that influence the prediction. However, they do not necessarily make the model itself transparent, since the explanation is produced only after training.

For this reason, the present thesis follows an interpretable-by-design direction. Instead of applying an explanation method only after training a black-box model, the architecture is chosen so that interpretability is part of the model structure. Prototype-based models make it possible to associate the learned representation with representative image patterns. In MRI-based retrieval, this provides a more direct link between the input scan, the learned embedding, and the visual content used for similarity estimation. This choice is consistent with recent prototype-based approaches in neuroimaging, where explanations are based on learned prototypical regions rather than only on external post-hoc visualizations.

2.5.2 Patch-Based Intuitive Prototypes Network: PIP-Net

PIP-Net (Patch-Based Intuitive Prototypes Network) is an interpretable image classification model that learns prototypical parts in a self-supervised manner and aligns well with human visual reasoning [28]. Extending this approach, a recent study [29] applied the framework to Alzheimer's disease, introducing a prototype-based deep learning architecture for volumetric medical images. Prototype-based models learn representative anatomical patterns directly from MRI scans and associate them with diagnostic classes. These learned prototypes correspond to meaningful brain regions, such as areas affected by neurodegeneration, and therefore provide clinically interpretable descriptors of patient images [29].

PIP-Net3D consists of a 3D CNN for feature extraction, a global max-pooling operation that produces prototype presence scores, and a linear classification layer that links prototypes to diagnostic classes. In contrast to 2D approaches that rely on selecting representative slices, PIP-Net3D operates directly on full 3D MRI volumes using 3D convolutional layers. This allows the model to capture volumetric structural patterns across multiple slices and automatically identify relevant anatomical regions without manual slice selection [29].

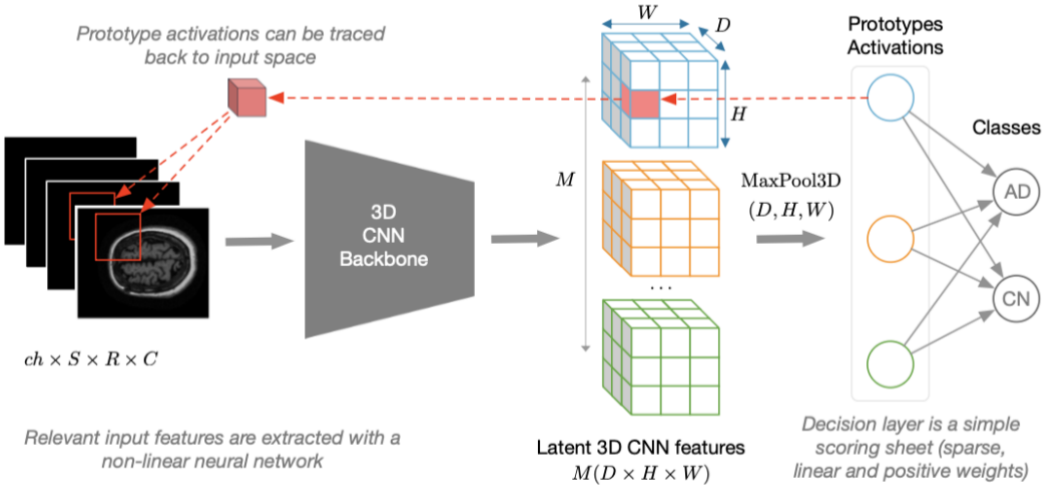

Figure 3: *Overview of PiPNet3D [29]*

After training, PIP-Net selects only the relevant prototypes. A prototype is kept if it contributes to the classification layer (weight > 10−3) and if it is activated in the training data with similarity > 0.1. This pruning step reduces the initial set of hundreds of prototypes to a small number of meaningful prototypes.

To analyze the consistency of prototype localization, an additional metric called Prototypes Consistency Localization was introduced. This measure evaluates whether the same prototype is activated in similar brain regions across different images.

This approach has the potential to improve the transparency and interpretability of deep learning-based neuroimaging systems, facilitating clinical validation of retrieved cases and supporting more explainable decision-making in dementia analysis.

The interpretability of PIP-Net3D arises from its prototype-based decision mechanism. Although the model is initially trained with a large set of prototypes, only a small subset remains active after training and pruning, typically between four and eleven prototypes. Each prototype tends to be associated with a single class, providing class-specific evidence that contributes to

the final decision. Moreover, most prototypes are consistently localized in similar anatomical regions across different MRI scans, which supports their interpretability [29].

Local explanations are also compact, with predictions typically relying on only a few prototypes. The study further reports that models with purer prototypes achieve higher classification performance. Finally, expert evaluation by radiologists confirmed that the majority of prototypes correspond to anatomically coherent regions and exhibit patterns consistent with known biomarkers of Alzheimer's disease, supporting the clinical plausibility of the model's explanations.

3. Methodology

The methodology was structured in three main steps. First, a CNN baseline was developed to predict NeuroCBIR-like embeddings directly from minimally preprocessed T1-weighted MRI scans, reducing the dependence on computationally expensive preprocessing pipelines. Second, the same dataset was used to train a PIP-Net3D-based model, adapted to produce embedding representations while introducing prototype-based interpretability. Finally, both approaches were evaluated and compared in a CBIR setting by analyzing the quality of the predicted embedding space, the similarity ranking of retrieved scans, and retrieval metrics such as mAP@k and success@k. This workflow allowed the study to assess whether interpretable models can reproduce meaningful MRI retrieval representations while improving transparency.

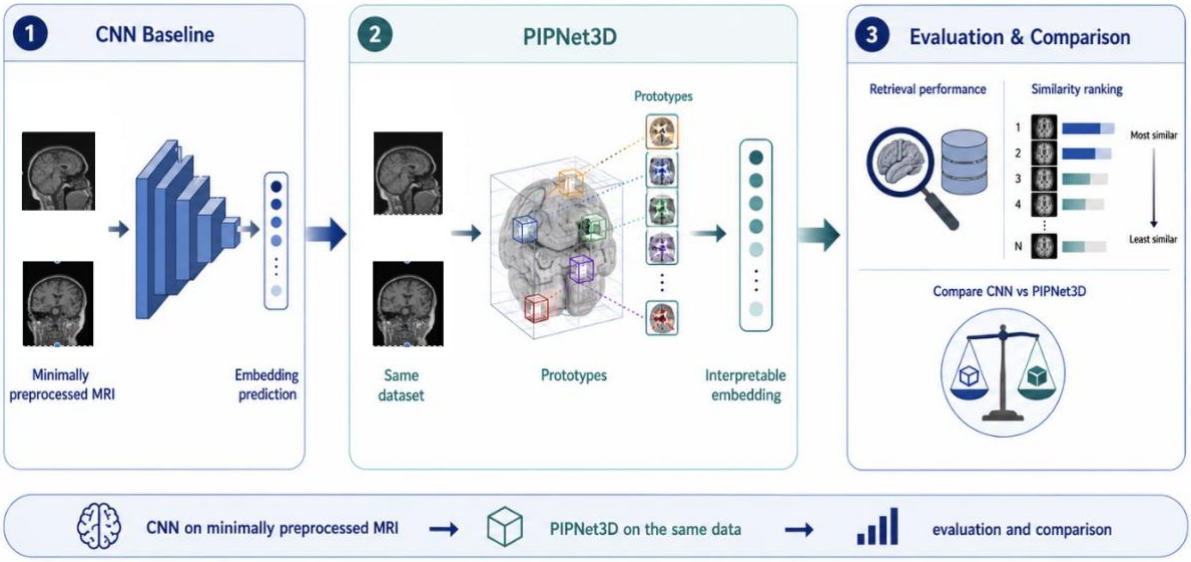

Figure 4: *Overview of the main methods implemented in this project, including the CNN baseline, the PIP-Net-based model, and the retrieval evaluation pipeline.*

3.1 Dataset

This study is based on secondary data analysis, using structural brain MRI scans and associated metadata from the OASIS-3 dataset [30], a publicly available longitudinal neuroimaging cohort focused on aging and Alzheimer's disease. To handle high-dimensional 3D MRI data efficiently, images were stored in chunked files that reduced memory bottlenecks during training.

The dataset consists of 2,133 MRI scans corresponding to 1,302 unique subjects, indicating the presence of longitudinal acquisitions. Each scan is uniquely identified by a GUID, with no duplicates, ensuring data consistency.

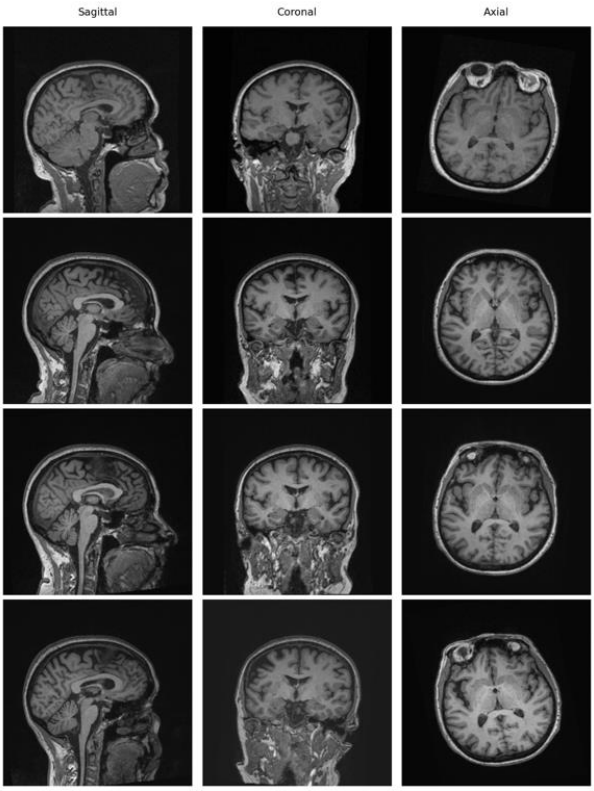

Figure 5: *Longitudinal scans for subject OAS30959*

The acquisition characteristics were derived from the dataset metadata. Specifically, the dataset includes a single scanner manufacturer, two magnetic field strengths (1.5T, 3T) and six scanner models, indicating limited variability in acquisition conditions. Each data point is enriched with clinical, demographic, and longitudinal metadata. Specifically, clinical labels include cognitively normal (CN), mild cognitive impairment, and Alzheimer's disease. Longitudinal information is available through subject identifiers and acquisition timepoints. Figure 6 shows the distribution of MRI acquisitions per subject in the OASIS-3 dataset. Most subjects have a single scan, while a substantial proportion underwent multiple MRI examinations over time. This longitudinal structure is fundamental for the retrieval evaluation, as only subjects with more than one acquisition can be used as queries, enabling the retrieval of follow-up scans from the same individual

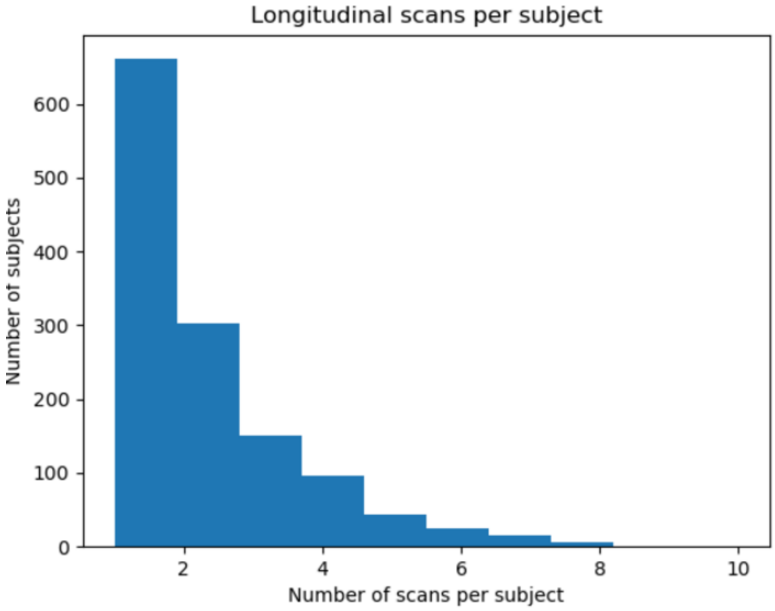

Figure 6: *Distribution of the number of MRI acquisitions per subject in the OASIS-3 dataset. Subjects with multiple acquisitions provide the longitudinal information required for the retrieval evaluation.*

All MRI volumes are standardized to a fixed input shape of (196, 196, 196), ensuring both consistency across samples and compatibility with 3D convolutional neural networks.

In this project, the focus is on using minimally processed MRI scans rather than heavily engineered features, in order to evaluate more clinically realistic pipelines.

3.1.1 External Validation Datasets

To assess the generalization of the proposed models beyond OASIS-3, an additional evaluation was performed on the external MIRIAD [32], SLIM [33] and AIBL [34] datasets. These datasets were not used during training, validation, or model selection. The trained CNN and PIP-Net models were applied without further fine-tuning to generate a 32-dimensional embedding for each MRI scan. This experiment was used to examine whether the predicted embedding spaces remained suitable for subject re-identification when applied to data from a different cohort.

The Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset was specifically developed for the analysis of longitudinal brain changes in Alzheimer's disease. It includes 46 patients with mild-to-moderate Alzheimer's disease and 23 non-demented controls, all older than 55 years. The dataset contains 708 volumetric T1-weighted MRI scans acquired by the same radiographer using the same 1.5 T GE scanner and acquisition protocol. Imaging sessions were scheduled at baseline and after 2, 6, 14, 26, 38, and 52 weeks, with additional scans obtained from a subset of participants after 18 and 24 months [32].

The Southwest University Longitudinal Imaging Multimodal (SLIM) dataset is a longitudinal neuroimaging study of healthy young adults. At the first time point, it included 580 participants with a mean age of 20.08 years and an age range of 17–27 years. A total of 240 and 228 participants were assessed at the second and third time points, respectively, while 121 participants completed all three sessions. Participants were screened to exclude neurological and psychiatric disorders. The dataset provides structural MRI, resting-state functional MRI, diffusion-weighted imaging, and behavioral measurements [33].

The Australian Imaging, Biomarkers and Lifestyle (AIBL) study is a longitudinal cohort designed to investigate biomarkers and risk factors associated with Alzheimer's disease. The original cohort included 1,112 participants over 60 years of age: 768 cognitively healthy controls, 133 individuals with mild cognitive impairment, and 211 patients with Alzheimer's disease. Participants underwent cognitive and clinical assessments, blood collection, and health and lifestyle questionnaires [34].

Unlike AIBL and MIRIAD, SLIM contains only healthy young adults. It therefore provides a substantially different population for assessing the generalization of models trained predominantly on older individuals.

3.2 CNN Baseline

In this initial stage of the project, a baseline model was developed to learn a mapping from minimally processed 3D T1-weighted MRI volumes to a predefined embedding space. The goal of this step was to provide a reference representation that could later be compared with more advanced approaches, such as the prototype-based PIP-Net model.

Subject-level metadata were loaded from a CSV file and used to match each MRI volume with its corresponding subject-level information. This matching was performed through the GUID identifiers shared between the metadata file, the NPZ image chunks, and the NeuroCBIR embedding file. No additional image preprocessing was applied at this stage. A cross-validation strategy was adopted, where three folds were used for training, one fold was used for validation, and the fifth fold, CV5, was kept as a held-out test set.

MRI volumes were paired with target embeddings generated by the NeuroCBIR framework, which served as supervisory signals in a regression setting. The proposed CNN was trained to approximate these embeddings, effectively learning to reproduce the NeuroCBIR representation from minimally processed MRI data.

3.2.1 Computational Resources

Experiments were conducted on Alvis, a National Academic Infrastructure for Supercomputing in Sweden (NAISS) resource dedicated to AI and Machine Learning research, clustered at Chalmers University of Technology. Training was performed using V100 NVIDIA GPUs.

3.2.2 Model Structure

The model architecture is a 3D convolutional encoder based on residual blocks, designed to regress NeuroCBIR embeddings from minimally processed MRI volumes. The input consisted of a single-channel 3D T1-weighted MRI volume with an initial spatial size of 196×196×196. The stem applied a strided 3D convolution to reduce the spatial resolution from $196^3\ to\ 98^3$ while increasing the number of channels to 16. This was followed by four convolutional stages with channel widths [16,32,64,128] and residual block depths [1,1,2,2]. Except for the first stage, each stage performed spatial downsampling before applying the residual blocks. The final feature representation was obtained through global average pooling, followed by dropout and a linear regression head that projected the representation into a 32-dimensional embedding space.

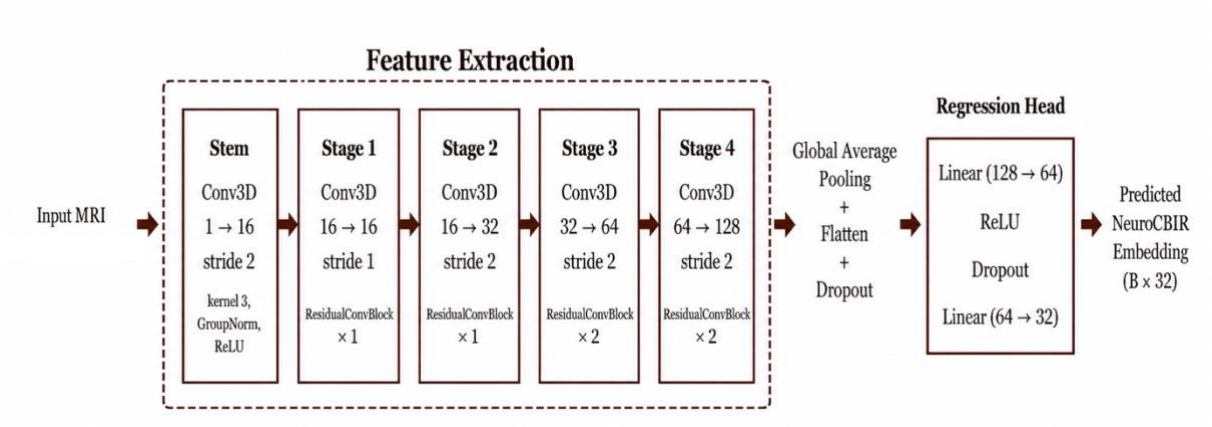

Figure 7: *CNN model Structure*

The CNN baseline was trained to predict a 32-dimensional NeuroCBIR embedding from each input MRI volume:

$$\hat{z} = f_\theta(x) \tag{2}$$

where x is the input MRI volume, f_θ is the CNN model, and $\hat{z} \in \mathbb{R}^{Bxd}$ is the predicted embedding tensor for a batch of size B, with embedding dimension d=32. The corresponding NeuroCBIR target embeddings are denoted as $z \in \mathbb{R}^{Bxd}$.

The model was trained using the AdamW optimizer with weight decay regularization to improve generalization. Half precision float16 training was employed to reduce memory consumption and accelerate computation, which is particularly advantageous when processing high-resolution 3D MRI volumes.

The training objective was defined as a combination of mean absolute error (MAE) and a correlation-based regularization term between the predicted and target embeddings.

The mean absolute error was computed as:

$$\mathcal{L}_{MAE} = \frac{1}{B\ d}\sum_{i=1}^{B}\sum_{j=1}^{d}\left|\hat{z}_{ij} - z_{ij}\right| \tag{3}$$

where $\hat{z}_{ij}\ is\ the\ j-th$ predicted embedding component of sample i, and z_{ij} is the corresponding NeuroCBIR target component.

In addition, a correlation coefficient was computed between the predicted and target embedding tensors after flattening all embedding values within the batch and mean-centering them. This auxiliary term was introduced to discourage degenerate solutions, such as predictions collapsing towards mean embedding values:

$$r = \frac{\sum_{i=1}^{B}\sum_{j=1}^{d}(\widehat{z_{ij}}-\bar{\hat{z}})(z_{ij}-\bar{z})}{\sqrt{\sum_{i=1}^{B}\sum_{j=1}^{d}(\widehat{z_{ij}}-\bar{\hat{z}})^2}\sqrt{\sum_{i=1}^{B}\sum_{j=1}^{d}(z_{ij}-\bar{z})^2}} \tag{4}$$

where B is the batch size, d= 32 is the embedding dimensionality, $\overline{\hat{z}}$ and $\bar{z}$ are the mean values of the predicted and target embeddings over the whole batch:

$$\bar{\hat{z}} = \frac{1}{Bd}\sum_{i=1}^{B}\sum_{j=1}^{d}\widehat{z_{ij}} \tag{5}$$

$$\bar{z} = \frac{1}{Bd}\sum_{i=1}^{B}\sum_{j=1}^{d}z_{ij} \tag{6}$$

The final loss was defined as the MAE minus the weighted correlation term:

$$\mathcal{L}_{total} = \mathcal{L}_{MAE} - \lambda_{corr} * r \tag{7}$$

where λ_{corr} = 0.05 and controls the contribution of the correlation term.

The negative sign before the correlation term ensured that minimizing the total loss encouraged higher correlation between the predicted and target embeddings. In this way, the CNN was trained through embedding distillation: it learned to approximate the retrieval-oriented NeuroCBIR embedding space from minimally processed MRI volumes, rather than directly optimizing a contrastive retrieval loss.

Model performance was monitored on the validation set using only MAE as the early stopping criterion, and the best-performing checkpoint was subsequently evaluated on the held-out test set.

This baseline serves as a reference for evaluating the added value of interpretability-driven and retrieval-based approaches introduced in the following sections.

Table 1: *Hyperparameter values*

Hyperparameter	***Value***
Optimiser	AdamW
Learning Rate	2×10^{-4}
Dropout	0.4
Weight Decay	1×10^{-4}
Loss	MAE + correlation loss
Maximum epochs	10 000
Early stopping patience	1000

3.2.3 Model Optimization

Starting from the model trained with the MAE + correlation loss described in the previous section, the learned weights were used as initialization for a subsequent fine-tuning phase in which the loss function was replaced with Smooth L1 loss. This second stage was introduced to refine the alignment between the predicted embedding and the NeuroCBIR target embedding, without restarting training from scratch.

Let $(\hat{z}ij)$ be the j-th predicted embedding component for subject i and *let* zij be the corresponding NeuroCBIR target embedding component. The prediction error is defined as:

$$e_{ij} = \hat{z}ij - zij \tag{8}$$

The Smooth L1 loss with parameter (β) is defined as:

$$\ell_{\beta(e_{ij})} = \begin{cases} 0.5\dfrac{e_{ij}^2}{\beta}, & if\ |e_{ij}| < \beta; \\ |e_{ij}| - 0.5\beta, & otherwise \end{cases} \tag{9}$$

In this work, (β = 0.1) was used. With this setting, embedding components with an absolute prediction error smaller than 0.1 are optimized using the quadratic part of the loss, while larger errors are treated with the linear L1 component.

The motivation for introducing Smooth L1 loss in this second phase is related to the behavior of the optimization near convergence. After the initial MAE + correlation training, the model already produces embeddings that are globally aligned with the NeuroCBIR space. At this stage, the objective shifts from coarse alignment to fine-grained element-wise refinement. Smooth L1 provides a stable transition between L2-like behavior for small errors, which encourages precise adjustments, and L1-like behavior for larger residuals, which prevents large gradients from dominating the update. This makes it particularly suitable for fine-tuning the embedding regression once the model is already close to the target representation.

3.3 PIP-Net for Embedding Regression

In this thesis, a modified version of PIP-Net3D was implemented to obtain interpretable MRI representations for content-based image retrieval. The original PIP-Net3D model was proposed as a prototype-based classifier for 3D MRI scans. In that formulation, the network extracts prototype activation maps from an input MRI volume and combines the resulting prototype scores through a sparse linear classification layer to predict diagnostic classes such as Alzheimer's disease or cognitively normal status.

In contrast, the model used in this work was adapted from a classification task to an embedding regression task. The objective was not to directly classify the MRI scan, but to predict a 32-dimensional NeuroCBIR embedding from a minimally processed T1-weighted MRI volume. Therefore, the original classification layer was removed and replaced by a linear embedding layer. This modification allows the network to preserve the prototype-based structure of PIP-Net, while producing an embedding that can be used for similarity-based image retrieval.

Let x_i be the input MRI scan and e_i be the corresponding target NeuroCBIR embedding:

$$x_i \in R^{C\times D\times H\times W}$$

$$e_i \in R^{32}$$

where C is the number of input channels and D, H, and W are the spatial dimensions of the MRI volume. The goal of the network is to learn a function that maps the input image to a predicted embedding:

$$\widehat{e_i} = F(x_i) \tag{10}$$

Where

$$\widehat{e_i} \in R^{32}$$

The predicted embedding $\widehat{e_i}$ is then trained to approximate the target NeuroCBIR embedding e_i. In this way, the model learns to reproduce the NeuroCBIR representation space while maintaining access to intermediate prototype activations that can be inspected for interpretability.

3.3.1 Dataset Construction

The training pipeline used MRI volumes stored in NPZ files and target NeuroCBIR embeddings stored in a parquet file. Each MRI scan was associated with a unique identifier, which was used to match the image volume with its corresponding 32-dimensional embedding. The dataset was divided into five folds. The fifth fold was kept as a fixed test set, while the first four folds were used for training and validation using a leave-one-fold-out strategy.
Two dataset classes were used. The first dataset, referred to as the single NPZ dataset, returns one MRI volume and its target embedding:

$$(x_i, e_i)$$

This dataset was used for validation, testing, projection, and embedding extraction, where deterministic evaluation is required. The second dataset, referred to as the two-augmentation dataset, returns two independently augmented versions of the same MRI volume together with the same target embedding:

$$\left(\widetilde{x_i^{(1)}}, \widetilde{x_i^{(2)}}, e_i\right)$$

where

$$\widetilde{x_i^{(1)}} = T_1(x_i), \qquad \widetilde{x_i^{(2)}} = T_2(x_i) \tag{11}$$

Here, T1 and T2 represent two random augmentation transformations applied to the same original MRI scan. The rationale is that small spatial or intensity transformations should not change the anatomical identity of the image. Therefore, the two augmented views should activate similar prototypes and should be mapped to the same NeuroCBIR embedding.

This strategy is important for prototype learning. If two slightly transformed versions of the same MRI scan are given to the network, the model is encouraged to learn prototypes that are stable under augmentation, rather than prototypes that depend on small irrelevant image variations.

3.3.2 Network Architecture

The proposed model is composed of five main blocks: the feature network, the add-on layer, the prototype activation maps, the pooling layer, and the embedding layer.

The feature network is a 3D CNN backbone that extracts latent features from the input MRI volume. The purpose of testing two backbones was to evaluate whether different feature extractors influence the quality of the learned prototype representations and the final embedding prediction. Both backbones proposed in the original PIP-Net3D framework were initially explored. Since preliminary ConvNeXt experiments produced substantially poorer results, the final experiments were conducted using the 3D ResNet-18 backbone.

Given an input volume x_i, the backbone extracts a latent feature representation:

$$h_i = f_\theta(x_i) \tag{12}$$

Where f_θ represents the 3D CNN backbone with parameters θ.

The latent features extracted by the backbone are passed to the add-on layer, which converts them into prototype activation maps.

The output of this layer is a tensor:

$$A_i \in R^{M \times d \times h \times w}$$

where M is the number of prototypes and d, h, and w are the spatial dimensions of the prototype activation maps.

Each channel of A_i corresponds to the activation map of one prototype. Therefore, the value:

$$A_{i,m,u}$$

represents the activation of prototype m at spatial location u, where:

$$u \in \Omega$$

and Ω is the set of all spatial locations in the feature map. In the implemented model, the add-on layer applies a softmax along the prototype dimension. This encourages each spatial location to be associated more strongly with a subset of prototypes:

$$A_{i,m,u} = \frac{\exp(q_{i,m,u})}{\sum_{r=1}^{M} \exp(q_{i,r,u})} \tag{13}$$

The pooling layer converts the spatial prototype activation maps into a compact prototype presence vector. For each prototype, a 3D max-pooling operation selects the maximum activation across all spatial locations:

$$p_{i,m} = \max_{u \in \Omega} A_{i,m,u} \tag{14}$$

The resulting prototype vector is:

$$p_i = [p_{i,1}, p_{i,2}, \dots, p_{i,M}]$$

With

$$p_i \in R^M$$

Each element $p_{i,m}$ represents how strongly prototype m is present in the MRI scan.

In the original PIP-Net3D model, the pooled prototype vector is passed to a sparse linear classification layer. In this thesis, that layer was replaced by a linear embedding layer. This layer maps the prototype presence vector to a 32-dimensional embedding:

$$\widehat{e_i} = W_{\text{emb}} p_i + b_{\text{emb}} \tag{15}$$

Where:

$$W_{\text{emb}} \in R^{32 \times M}$$

And

$$b_{\text{emb}} \in R^{32}$$

The final output of the model is therefore:

$$\widehat{e_i} \in R^{32}$$

This embedding can then be used in the same way as NeuroCBIR embeddings for similarity-based retrieval.

3.3.3 Loss Functions

The model was trained using a combination of three losses: an alignment loss, a tanh loss, and an embedding regression loss based on mean absolute error. Each loss has a specific role in the training process.

The alignment loss encourages the prototype feature maps of two augmented versions of the same MRI scan to be similar. Given two augmented views $\widetilde{x_i^{(1)}}$ *and* $\widetilde{x_i^{(2)}}$ the model extracts two prototype feature tensors.

The purpose of this loss is to make the prototype activations invariant to small augmentations of the same MRI scan. In other words, if the same brain is slightly rotated, translated, or perturbed with noise, the activated prototypes should remain consistent.

The tanh loss encourages the model to use the available prototypes. Without this term, some prototypes could remain inactive during training. This loss promotes prototype activation across the batch and helps the network avoid degenerate solutions in which only a small subset of prototypes is used.

The embedding regression loss compares the predicted embedding $\widehat{e_i}$ with the target NeuroCBIR embedding e_i. Since the final embedding has 32 dimensions, the mean absolute error is computed across the embedding coordinates

$$\mathcal{L}_{MAE} = \frac{1}{32}\sum_{j=1}^{32}\left|\widehat{e_{i,J}} - e_{i,j}\right| \tag{16}$$

Since each training sample is represented by two augmented views, the target embedding is duplicated: $e_i^{(1)} = e_i^{(2)} = e_i$.

Therefore, both augmented views are forced to predict the same NeuroCBIR embedding:

$$\mathcal{L}_{emb} = \frac{1}{2}\left[\frac{1}{32}\sum_{j=1}^{32}\left|\widehat{e_{i,J}^{(1)}} - e_{i,j}\right| + \frac{1}{32}\sum_{j=1}^{32}\left|\widehat{e_{i,J}^{(2)}} - e_{i,j}\right|\right] \tag{17}$$

This term directly optimizes the model to reproduce the target NeuroCBIR embedding space.

In addition to the regression objective, an in-batch contrastive loss is applied to the predicted embeddings. The embeddings from the two views are first L2-normalized:

$$z_i^{(v)} = \frac{\widehat{e_i^{(v)}}}{|\widehat{e_i^{(v)}}|_2}, \quad v \in \{1,2\} \tag{18}$$

The two augmented views of the same scan form a positive pair, whereas embeddings associated with different scans in the batch are treated as negative pairs.

This term encourages the two augmented views of the same MRI scan to produce similar embeddings while separating them from the embeddings of other scans within the batch.

The complete training objective is defined as:

$$\mathcal{L}_{tot} = \lambda_{\text{align}}\mathcal{L}_{align} + \lambda_{tanh}\mathcal{L}_{tanh} + \lambda_{\text{MAE}}\mathcal{L}_{emb} + \lambda_{\text{cont}}\mathcal{L}_{cont} \tag{19}$$

where λ_{align}, λ_{tanh}, λ_{MAE} and λ_{cont} control the contribution of each loss term.
The alignment and tanh terms guide the learning of stable and meaningful prototypes, while the MAE and contrastive terms ensure that the final output approximates the NeuroCBIR embedding.

3.3.4 Training Procedure

During the first phase, only the prototype learning objectives were used. The embedding layer was frozen, and the model was optimized using the alignment loss and the tanh loss:

$$\mathcal{L}_{tot} = \lambda_{\text{align}}\mathcal{L}_{align} + \lambda_{tanh}\mathcal{L}_{tanh} \quad (20)$$

During this phase, the alignment weight increased progressively according to $\lambda_{\text{align}} = e/E$, where e is the current epoch and E is the total number of pretraining epochs. The purpose of prototype pretraining was to initialize the prototype representation before learning the embedding regression task. The alignment term encourages corresponding patches from two augmented views of the same MRI scan to produce consistent prototype activations, while the tanh term promotes prototype activation across the training batch.
After prototype pretraining, the network parameters were frozen and only the embedding head was optimized for the first 10 training epochs. In this phase, the model learns how to map the already learned prototype presence vector p_i to the target NeuroCBIR embedding e_i.
The loss used in this phase is:

$$\mathcal{L}_{\text{train}} = \lambda_{\text{MAE}}\mathcal{L}_{emb} + \lambda_{\text{cont}}\mathcal{L}_{cont} \quad (21)$$

The embedding regression term encourages the predicted output to approximate the target NeuroCBIR embedding, whereas the contrastive term encourages the two augmented views of the same scan to produce similar embeddings and separates them from other scans in the batch. This stage provides an initial optimization of the embedding head before the prototype-producing layers are updated. After this initial head-training stage, the add-on layer and the trainable feature-extraction blocks were unfrozen. The backbone remained frozen until the predefined freezing period was completed and was subsequently unfrozen for end-to-end optimization.
This gradual unfreezing strategy allows the model to preserve the prototype representation learned during pretraining while progressively adapting it to the embedding regression task. In the final end-to-end stage, the backbone, add-on layer, and embedding head are optimized

jointly. The alignment and tanh terms continue to regularize the prototype representation, while the regression and contrastive terms encourage the output to reproduce the structure of the NeuroCBIR embedding space.

Overall, the proposed model reformulates PIP-Net3D from an interpretable classification network into a prototype-based embedding regression model. Prototype activations serve as an intermediate representation from which a 32-dimensional NeuroCBIR-like embedding is predicted. By operating on minimally processed T1-weighted MRI scans, the approach investigates whether the NeuroCBIR embedding space can be approximated with reduced dependence on computationally intensive preprocessing. At the same time, the spatial prototype activations provide a mechanism for examining which image regions contribute to the predicted representation and, consequently, to similarity-based retrieval.

3.4 Pretrained Brain MRI Model

An additional experiment was conducted to assess whether large scale pretraining on heterogeneous brain MRI data could improve the prediction of NeuroCBIR embeddings. For this purpose, the encoder of FOMO2JOMO, the highest ranked model in the FOMO25 Method Track, was used as a pretrained backbone. The original model was developed as a multimodal three-dimensional U-Net based variational autoencoder and was pretrained on the FOMO60K dataset using self-supervised learning [31].

In the present work, only the convolutional encoder and its pretrained weights were retained. The remaining components of the original architecture were replaced by a regression head designed to produce the same 32 dimensional embeddings used by NeuroCBIR. The complete network was then fine-tuned on the OASIS-3 training data, allowing the pretrained image representations to adapt to the embedding regression task. Model optimization combined a regression term with a cosine similarity term, while the best checkpoint was selected according to its performance on the validation partition. Data augmentation was applied during training to improve robustness to variations in image geometry and intensity.

3.5 Evaluation Metrics

To evaluate the quality of the learned embedding space, a Content-Based Image Retrieval (CBIR) evaluation was performed. Given a query MRI scan, the system retrieves the most similar scans from the reference dataset by comparing their predicted embeddings using cosine

similarity. This follows the NeuroCBIR evaluation protocol, where retrieval is formulated as a subject re-identification task: a retrieved scan is considered relevant if it belongs to the same subject as the query scan. This setting is motivated by the fact that longitudinal scans of the same subject share fine-grained anatomical patterns, and therefore a meaningful embedding space should place them close to each other. NeuroCBIR also evaluates retrieval by ranking reference embeddings according to pairwise cosine similarity and retrieving the top-k most similar scans.

The precision at k for each query was computed as:

$$\text{Prec@}k_i = \frac{\left|\text{rel}_\text{i}^{(\text{k})}\right|}{k} \tag{22}$$

where $\left|\text{rel}_\text{i}^{(\text{k})}\right|$ denotes the number of relevant images among the top-k retrieved results for query i.

The mean Average Precision at k was then obtained by averaging the precision values across all evaluated queries:

$$\text{mAP@}k = \frac{1}{N_{\text{queries}}} \sum_{i=1}^{N_{\text{queries}}} Prec\,@k_i \tag{23}$$

where N_{queries} is the total number of evaluated queries.

The mean Success at k was computed as:

$$\text{mS@}k = \frac{1}{N_{\text{queries}}} \sum_{i=1}^{N_{\text{queries}}} s_i \tag{24}$$

where $s_i = 1$ if query i has at least one relevant image in the top-k results, and $s_i = 0$ otherwise. In this work, a retrieval was considered successful when at least one of the retrieved images belonged to the same subject as the query image. Thus, mAP@k measures the average proportion of correct same-subject images retrieved among the top-k results, while mS@k measures how often at least one correct same-subject image is retrieved within the top-k.

3.5.1 Embedding-space Similarity

The embedding space was evaluated by testing whether the ranking produced by the embeddings was consistent with the structural similarity between MRI scans, using Spearman's rank correlation coefficient. For this purpose, a precomputed file containing multi-scale structural similarity index (MS-SSIM) values for selected pairs of scans was used [6].

This metric evaluates whether the embedding space is consistent with the structural similarity of the original MRI scans. In practice, images that are close in the embedding space should also appear similar according to MS-SSIM. Spearman's rank correlation is used to compare the ranking produced by the embeddings with the ranking suggested by MS-SSIM [6]. Since lower ranks indicate higher embedding similarity, better agreement corresponds to more negative correlation values.

Spearman's correlation is then computed as:

$$\rho_{\text{MS-SSIM}} = 1 - \frac{6 \sum_{(i,j)} \left(rank(s_{ij}) - rank(s_{\text{MS-SSIM},ij}) \right)^2}{(n^2 - 1)} \tag{25}$$

where n is the number of evaluated image pairs.

4. Results

This chapter reports the results of the experiments performed to evaluate whether minimally processed MRI scans can be used to reproduce NeuroCBIR embeddings. The models are assessed at two levels. First, the predicted embeddings are used in a content-based image retrieval setting, where performance is measured using mean Average Precision and mean Success. Second, the organization of the predicted embedding space is assessed by comparing embedding-based rankings with structural similarity measured by MS-SSIM using Spearman correlation.

The CNN regressor is first used as a baseline model. Then, the optimized PIP-Net-based architecture is evaluated and compared against the CNN baseline. Finally, qualitative results are reported to analyse whether the prototype-based model provides interpretable information about the image regions contributing to the embedding prediction.

4.1 Experimental Overview

All training and internal evaluation experiments were conducted on the OASIS-3 dataset, including scans from all available diagnostic classes in the selected cohort: cognitively normal subjects, subjects with mild cognitive impairment, and subjects with Alzheimer's disease. The target embeddings correspond to the NeuroCBIR representations previously extracted and stored in the projected embeddings file.

The dataset is divided into predefined folds. In the experiments reported in this section, fold 1 is used as the validation set, while the remaining training folds are used for model optimization. Fold 5 is kept as an independent test set and is used only for the final evaluation. Generalization was further evaluated on three external datasets: AIBL, MIRIAD, and SLIM.

4.2 CNN Baseline Results

4.2.1 Retrieval Metrics

After training, the best-performing checkpoint was evaluated on the independent test set (CV5). Table 2 reports the corresponding retrieval performance for fold 1.

Table 2: *Retrieval performance comparison between the original NeuroCBIR embeddings and the CNN-predicted embeddings for fold 1*

	mAP@1	**mAP@3**	**mAP@5**	**mS@1**	**mS@3**	**mS@5**
CNN-predicted embeddings	0.81	0.77	0.71	0.81	0.93	0.95
Original NeuroCBIR embeddings	0.94	0.96	0.95	0.94	0.99	0.99

Overall, this comparison provides a direct evaluation of whether the CNN baseline preserves the retrieval behaviour of NeuroCBIR when embeddings are predicted from minimally processed MRI scans.

4.2.2 Qualitative Retrieval Examples

To complement the quantitative retrieval metrics, qualitative examples were inspected to verify whether the predicted embedding space retrieved visually and anatomically plausible MRI scans. Figure 8 shows three representative query scans and their top-ranked retrieved cases.

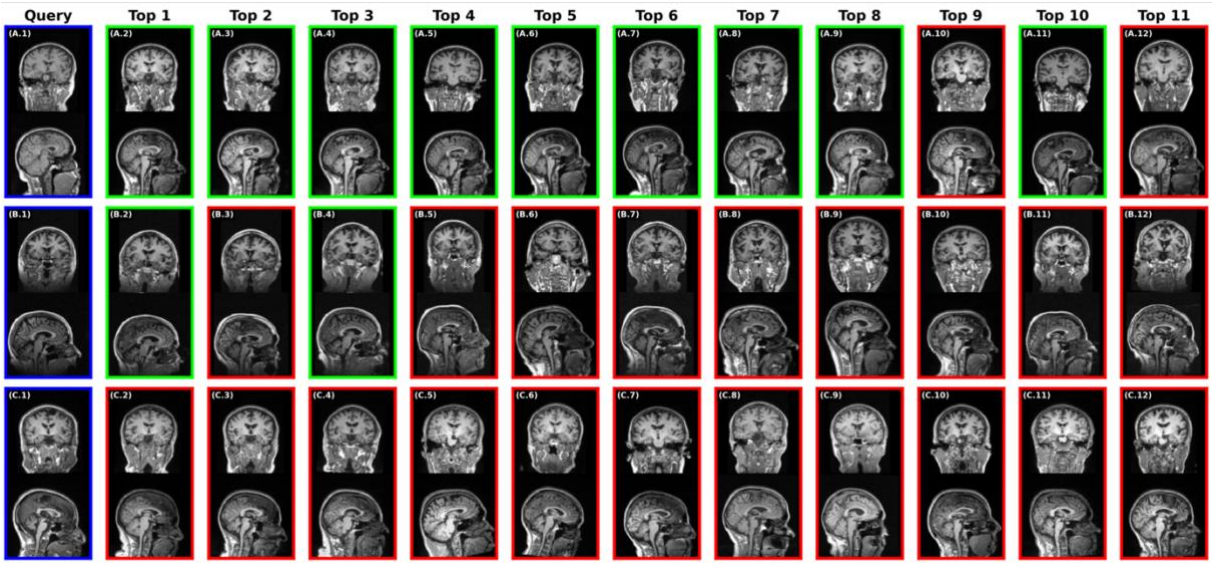

Figure 8: Qualitative retrieval examples for T1-weighted brain MRI scans. For each row, the first column shows the query scan, followed by the retrieved scans ranked from Top 1 to Top 11. A.1 corresponds to a subject with 10 available MRI scans, B.1 to a subject with 5 scans, and C.1 to a subject with 3 scans. The blue border marks the query, green borders indicate correct matches from the same subject, and red borders indicate scans from different subjects retrieved as visually similar.

4.2.3 Embedding Space Similarity

In addition to the retrieval metrics, the predicted embedding space was evaluated using the Spearman correlation with MS-SSIM. This analysis was used to assess whether images that are structurally similar according to MS-SSIM are also ranked as similar in the predicted embedding space. Therefore, this metric provides complementary information to mAP and mean Success, since it evaluates the consistency of the embedding space beyond same-subject retrieval.

The MS-SSIM correlation obtained with the CNN-predicted embeddings was compared with the value obtained using the original NeuroCBIR embeddings. This comparison allows us to assess whether the CNN model preserves not only the subject-level retrieval behavior, but also the broader structural similarity organization of the target embedding space.

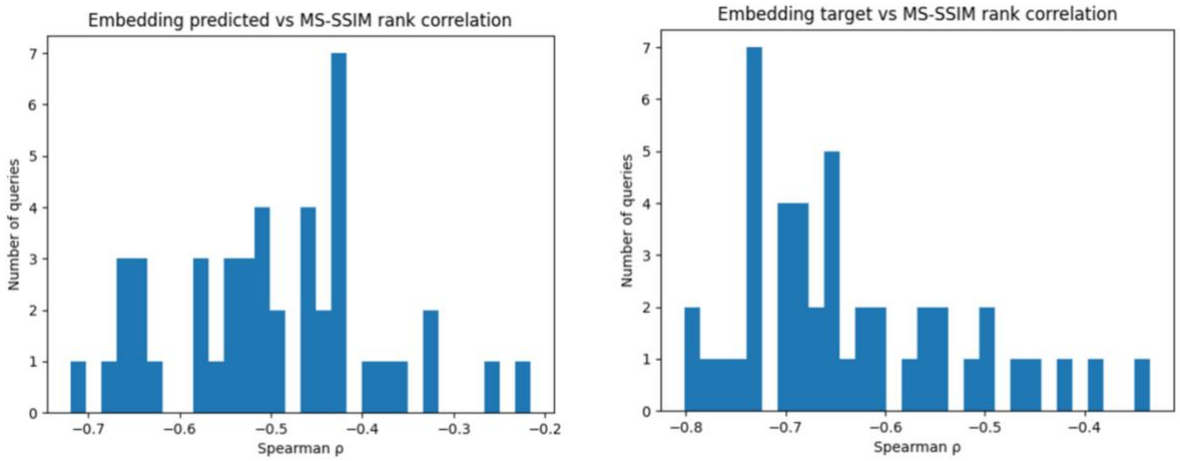

Figure 9: *Embedding space similarity graphs. The left panel refers to the embeddings predicted by the CNN model, while the right panel refers to the target NeuroCBIR embeddings. The plots compare embedding-based retrieval rank with MS-SSIM values*

Table 3 reports the distribution of Spearman correlation values between embedding-based retrieval rank and MS-SSIM for the target NeuroCBIR embeddings and the CNN-predicted embeddings.

Table 3: *Embedding space similarity analysis using MS-SSIM.*

Embedding source	**Mean ρ**	**Median ρ**	**Std**	**Min**	**Max**	**Evaluated queries**
Target NeuroCBIR embeddings	-0.637	-0.661	0.111	-0.802	-0.334	45
Predicted embeddings CNN	-0.498	-0.502	0.115	-0.720	-0.216	45

4.3 PIP-Net3D Results

4.3.1 Training Behaviour

The PIP-Net-based embedding regressor was trained using a 3D ResNet-18 backbone pretrained on Kinetics-400. Both backbone architectures proposed in the original PIP-Net3D framework were initially explored, including the ConvNeXt-Tiny variant pretrained on the STOIC medical imaging dataset. However, preliminary experiments with ConvNeXt produced substantially poorer embedding-regression and retrieval performance. This observation was consistent with the original PIP-Net3D study, in which the ResNet-based model also achieved higher predictive performance than the ConvNeXt-based variant, although that comparison was conducted for AD classification rather than embedding regression. Therefore, the subsequent experiments and analyses focused on the ResNet-18 backbone.

The model was not trained entirely from scratch. The 3D ResNet-18 feature extractor was initialized with pretrained weights, while the layers specific to the embedding-regression task were initialized from scratch. In particular, the original classification layer of PIP-Net3D was removed and replaced by a linear embedding head mapping the prototype presence vector to the 32-dimensional NeuroCBIR embedding as described in Section 3.3. The implemented model therefore preserved the prototype-based structure of PIP-Net but reformulated the final prediction task from classification to embedding regression. Figure 10 shows the training and validation loss curves used to monitor the optimization process and to assess the convergence behavior of the model.

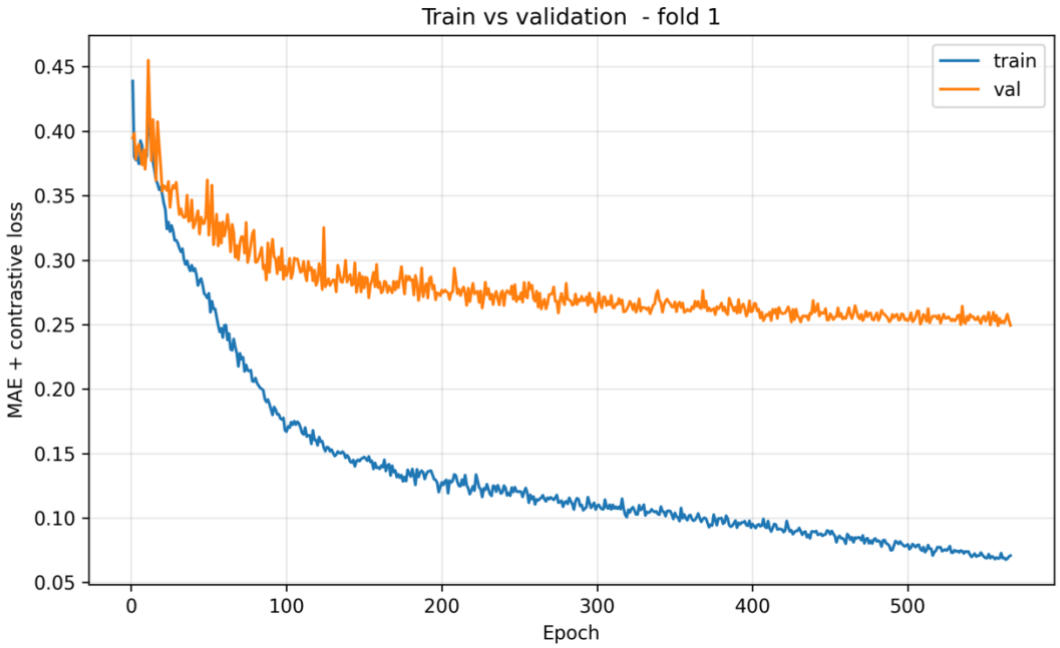

***Figure 10**: Training and validation of the PIP-Net-based embedding regressor on fold 1.*

4.3.2 Retrieval Metrics

After training, the best PIP-Net checkpoint was used to extract the predicted embeddings, which were then evaluated in the same CBIR setting used for the CNN baseline. Retrieval was performed using cosine similarity, and a retrieved scan was considered relevant when it belonged to the same subject as the query scan. Table 4 reports the retrieval performance of the PIP-Net-predicted embeddings (using fold 1 as validation) and compares it with the original NeuroCBIR embeddings.

***Table 4:** Retrieval performance comparison between the original NeuroCBIR embeddings and the PIP-Net-predicted embeddings*

Embedding source	mAP@1	mAP@3	mAP@5	mS@1	mS@3	mS@5
Original NeuroCBIR embeddings	0.940	0.960	0.950	0.940	0.990	0.990
PIP-Net predicted embeddings	0.869	0.833	0.792	0.869	0.951	0.951

4.3.3 Qualitative Retrieval Examples

To complement the quantitative retrieval metrics, qualitative examples were inspected to verify whether the predicted embedding space retrieved visually and anatomically plausible MRI scans. Figure 11 shows three representative query scans and their top-ranked retrieved cases.

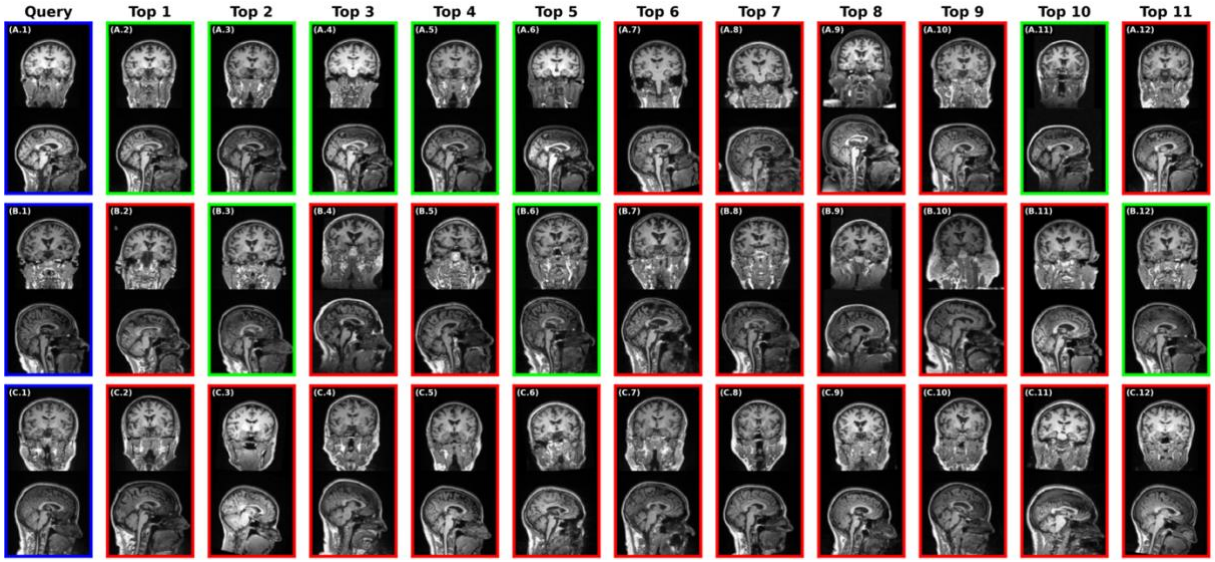

Figure 11: *Qualitative retrieval examples for T1-weighted brain MRI scans. For each row, the first column shows the query scan, followed by the retrieved scans ranked from Top 1 to Top 11. A.1 corresponds to a subject with 7 available MRI scans, B.1 to a subject with 7 scans, and C.1 to a subject with 4 scans. The blue border marks the query, green borders indicate correct matches from the same subject, and red borders indicate scans from different subjects retrieved as visually similar.*

4.3.4 Embedding Space Similarity

As described in Section 4.2.3, the Spearman correlation between embedding-based similarity and MS-SSIM was computed for the PIP-Net predicted embeddings and compared with the corresponding correlation obtained using the original NeuroCBIR embeddings.

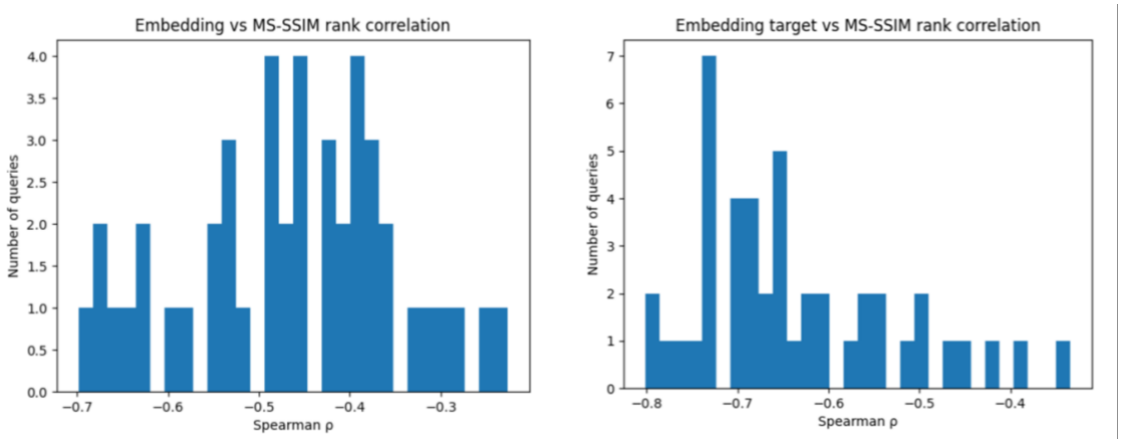

Figure 12: *Embedding space similarity graphs. The left panel refers to the embeddings predicted by the PIP-Net model, while the right panel refers to the target NeuroCBIR embeddings. The plots compare embedding-based retrieval rank with MS-SSIM values*

Table 5: *Embedding space similarity analysis using MS-SSIM.*

Embedding source	**Mean ρ**	**Median ρ**	**Std**	**Min**	**Max**	**Evaluated queries**
Target NeuroCBIR embeddings	-0.637	-0.661	0.111	-0.802	-0.334	45
Predicted embeddings PIP-Net	-0.463	-0.453	0.116	-0.698	-0.225	45

4.3.5 Prototype Visualization

Prototypes can be understood as local patterns that the network learns from the training images. When an MRI scan is processed, each prototype produces a three-dimensional activation map showing where a similar pattern occurs in the volume. The location of the strongest response can be mapped back to the original scan. In the examples shown in Figure 13, the white bounding box marks the image region associated with this maximum activation.

Their function in the proposed regression model is different from that in the original classification-based PIP-Net3D. In the classification setting, prototype activations are connected to the diagnostic classes through a sparse linear layer. Only a limited number of prototypes therefore contribute to the predicted class [28-29]. Here, the classification layer was replaced by a linear mapping from the prototype activations to the 32 dimensions of the NeuroCBIR embedding. The predicted embedding is consequently formed from the combined contribution of several prototypes, rather than from a small set of class-specific patterns.

To examine which image patterns had the greatest influence on the predicted representations, the prototypes were ranked according to the magnitude of their contribution to the embedding. The 20 prototypes with the largest contributions were extracted and visually inspected. Twelve of them were mainly associated with identifiable brain structures or regions containing relevant cerebral anatomy. The remaining eight responded near the limits of the brain, including cortical boundaries, background areas, and the interface between the head and the surrounding background.

This analysis also revealed a limitation of working with minimally preprocessed images. The MRI volumes were not registered to a shared anatomical template, meaning that corresponding structures were not necessarily located at the same coordinates in every subject. Differences in head position, orientation, brain size, and field of view make it difficult to establish whether a prototype consistently identifies the same anatomical region across the dataset. Activations close to the image boundaries may also reflect positioning or acquisition characteristics rather than cerebral anatomy. For these reasons, the visualizations provide a qualitative indication of the patterns involved in the construction of the embedding, but they should not be interpreted as precise anatomical localizations.

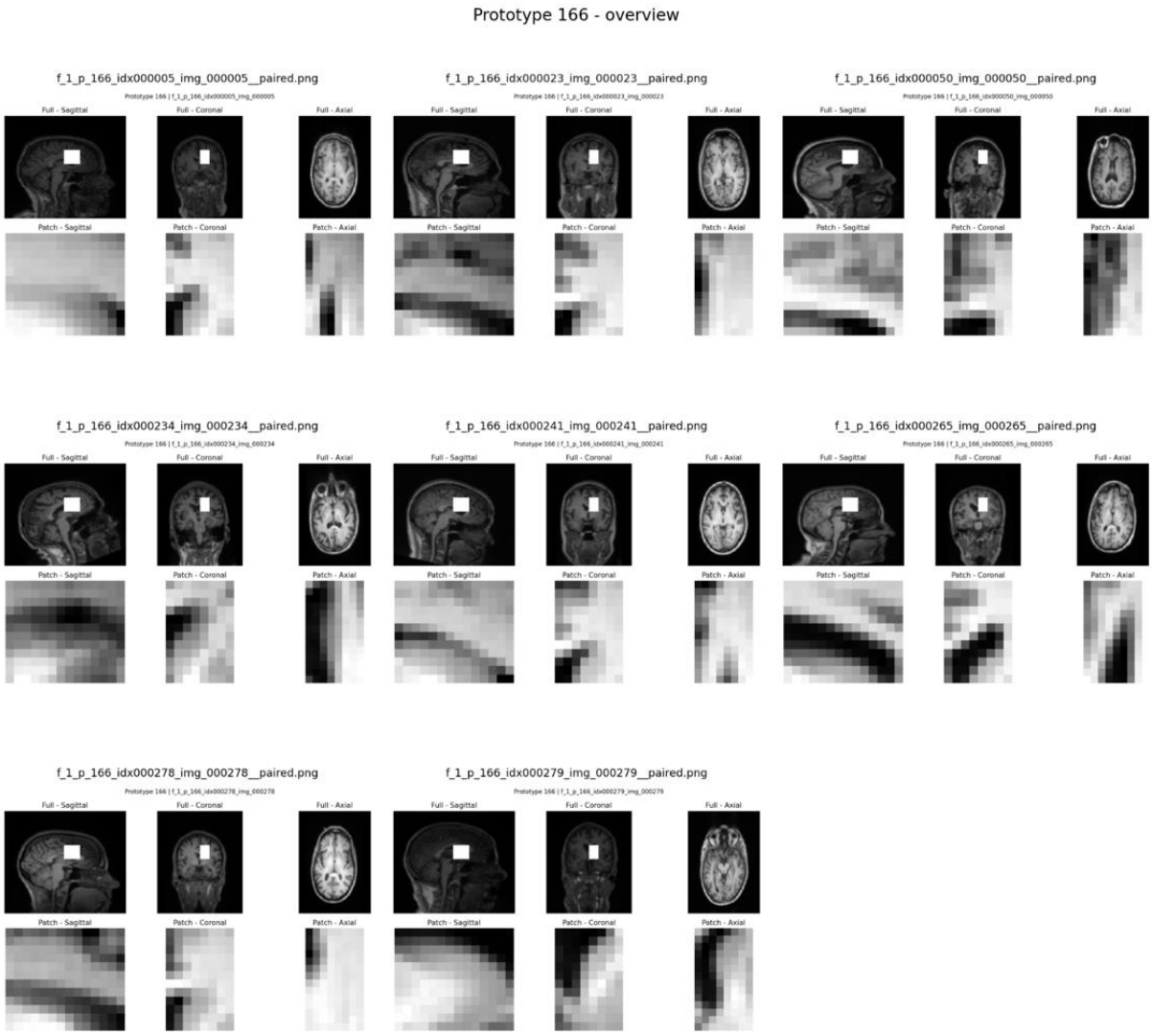

Figure 13: *Prototype 166 activation across different T1-weighted MRI scans. White boxes indicate the strongest activation, with the corresponding patches shown below.*

4.4 Cross-validation Results

Both for the CNN baseline and PIP-Net model, Folds 1-4 were used in turn as validation sets, while Fold 5 remained fixed for testing. In each run, the model was trained on the other three folds. Tables 6 and 7 report the overall OASIS-3 mean Success values obtained for each validation-fold configuration.

Table 6: *Mean Success across the four validation folds on CNN*

Validation fold	**mS@1**	**mS@3**	**mS@5**
Fold 1	0.806	0.925	0.946
Fold 2	0.566	0.663	0.655
Fold 3	0.430	0.591	0.577
Fold 4	0.647	0.828	0.815
Mean ± SD	0.613 ± 0.157	0.751 ± 0.152	0.748 ± 0.165

Table 7: *Mean Success across the four validation folds on PIP-Net*

Validation fold	**mS@1**	**mS@3**	**mS@5**
Fold 1	0.869	0.951	0.951
Fold 2	0.873	0.948	0.953
Fold 3	0.857	0.947	0.957
Fold 4	0.862	0.951	0.950
Mean ± SD	0.865 ± 0.0071	0.949 ± 0.0021	0.953 ± 0.0031

4.5 External Validation

External validation was performed on the MIRIAD, AIBL, and SLIM datasets without additional fine-tuning and using the models trained with Fold 1 as validation, which were also used in the main CNN–PIP-Net comparison.

As reported in Figure 14 and Table 8 , the predicted embeddings achieved the highest retrieval performance on MIRIAD, followed by AIBL, while lower results were obtained on SLIM.

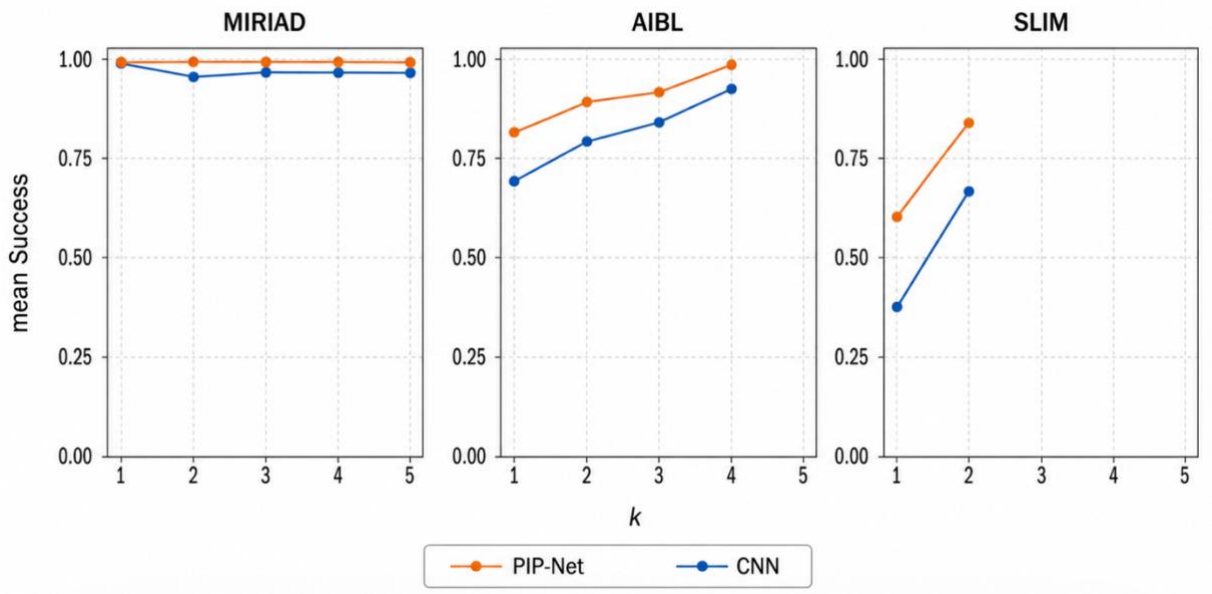

Figure 14: *Mean Success at k on the external datasets (MIRIAD, AIBL, and SLIM) for CNN- and PIP-Net-predicted embeddings*

Table 8: *Mean Success at k for CNN and PIP-Net on the external datasets.*

Dataset	**Model**	**mS@1**	**mS@2**	**mS@3**	**mS@4**	**mS@5**
MIRIAD	CNN	**0.995**	0.973	0.984	0.984	0.984
MIRIAD	PIP-Net	0.994	**0.996**	**0.996**	**0.996**	**0.995**
AIBL	CNN	0.696	0.800	0.859	0.940	N/A
AIBL	PIP-Net	0.827	0.891	0.920	0.981	N/A
SLIM	CNN	0.381	0.661	N/A	N/A	N/A
SLIM	PIP-Net	0.604	0.847	N/A	N/A	N/A

N/A indicates that no queries had at least k relevant same-subject scans available in the reference database, and therefore the metric could not be computed for that value of k.

4.6 Overall Comparison Between Models

Table 9 reports the retrieval results obtained from the original NeuroCBIR embeddings and from the embeddings predicted by PIP-Net, CNN baseline and FOMO2JOMO-based model. The original representation achieved the best performance for every metric. PIP-Net gave slightly higher scores than the CNN, with the largest difference observed for mAP@1. For both predicted representations, the success rate remained high as k increased, reaching 0.951 for PIP-Net and 0.946 for the CNN at k=5. The sharper decrease in mAP indicates that the models often retrieved at least one correct same-subject scan but were less reliable in keeping all relevant scans among the highest-ranked results.

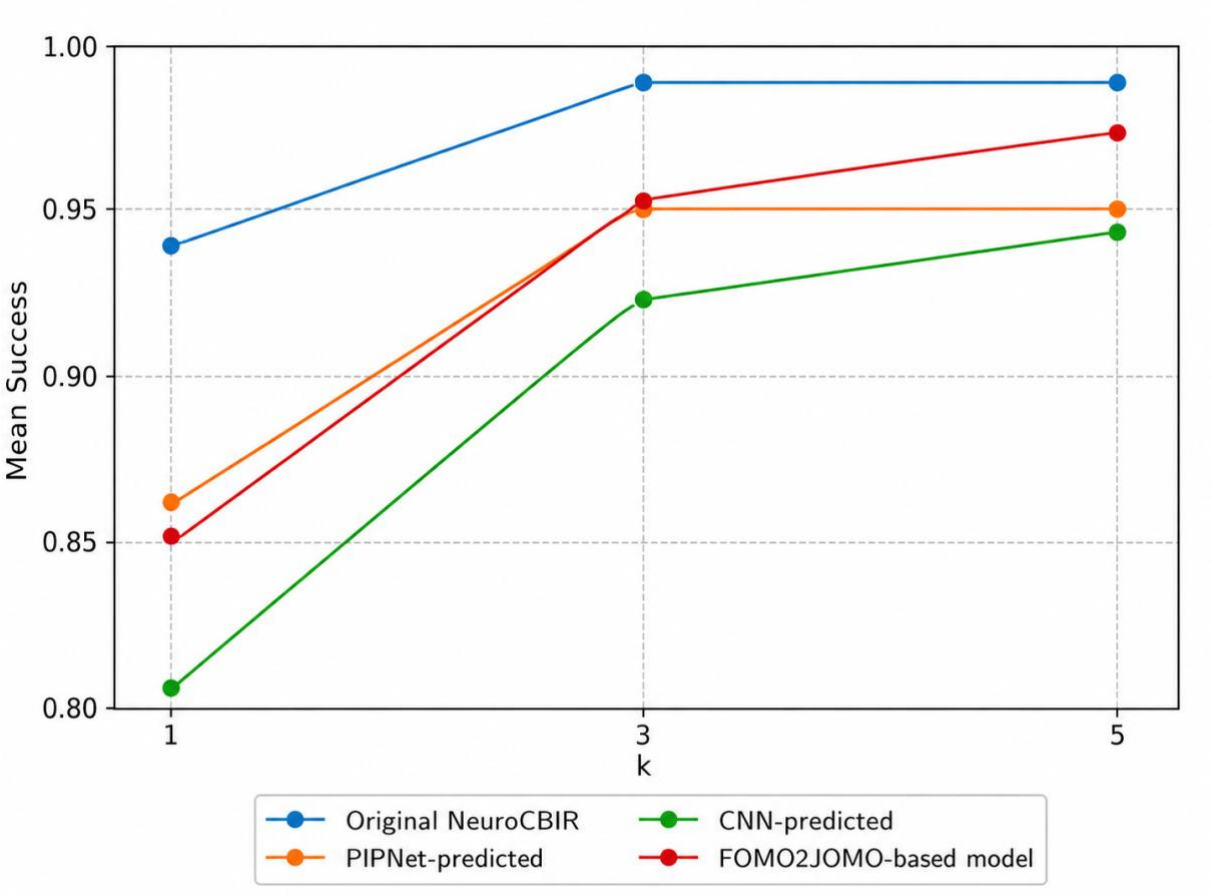

Figure 15: *Mean Success at k on OASIS-3 for the original NeuroCBIR, PIP-Net, CNN baseline, and FOMO2JOMO-based model, fold 1.*

Table 9: *Retrieval performance of the original NeuroCBIR, PIP-Net, CNN baseline and FOMO2JOMO- based model on the OASIS-3, fold 1.*

Embedding source	mAP@1	mAP@3	mAP@5	mS@1	mS@3	mS@5
Original NeuroCBIR	0.940	0.960	0.950	0.940	0.990	0.990
PIP-Net	0.869	0.833	0.792	0.869	0.951	0.951
CNN baseline	0.806	0.770	0.714	0.806	0.925	0.946
FOMO2JOMO-based model	0.859	0.821	0.783	0.859	0.954	0.974

To further inspect the retrieval behavior of the predicted embeddings, performance was evaluated across different subgroups.

Differences were observed across diagnostic groups. The highest performance was obtained for cognitively normal subjects, while lower values were observed for AD and MCI queries. However, this comparison should be interpreted with caution, because the number of AD and MCI queries was considerably smaller than the number of CN queries.

Table 10: *CNN and PIP-Net retrieval performance by diagnostic group for fold 1*

Disease	mS@3-CNN	mS@3-PIP-Net	mS@3-Neuro-CBIR
CN	0.929	0.955	0.995
AD	0.600	0.616	1.0
MCI	0.888	0.888	1.0

A larger difference was observed when grouping queries by scanner field strength. Images acquired at 3.0T achieved higher retrieval performance than those acquired at 1.5T. This may indicate a possible sensitivity of the predicted embedding space to acquisition-related differences, although the imbalance in the number of queries between the two groups should also be considered.

Table 11: *CNN and PIP-Net retrieval performance by scanner field strength for fold 1*

Field strength	mS@3-CNN	mS@3-PIP-Net	mS@3-Neuro-CBIR
1.5T	0.828	0.840	0.978
3.0T	0.934	0.962	0.997

5. Discussion

5.1 Main Findings

This thesis investigated whether NeuroCBIR embeddings [6] could be predicted directly from minimally processed T1-weighted MRI scans. Three deep learning approaches were considered: a conventional 3D CNN regressor, a prototype-based PIP-Net model adapted from classification to embedding regression, and an additional model based on a pretrained brain MRI encoder. The original NeuroCBIR embeddings were used as target representations, while the predicted embeddings were evaluated according to their ability to preserve subject-level retrieval and the broader similarity structure of the original embedding space.

Across the different validation-fold configurations, PIP-Net maintained strong retrieval performance on the fixed OASIS-3 and showed limited variation between folds. In the direct comparison performed using fold 1 as validation, the fine-tuned pretrained model achieved results close to those of PIP-Net. PIP-Net obtained slightly higher mean average precision, whereas the fine-tuned model reached marginally higher success rates for larger values of k. By contrast, the CNN baseline showed substantially greater variability across folds and lower overall retrieval performance.

Overall, the findings support the feasibility of approximating the NeuroCBIR embedding space from less extensively processed MRI scans. PIP-Net preserved robust retrieval performance while also providing spatial prototype activations that enabled inspection of the learned representation. The comparable performance of the fine-tuned model further suggests that large-scale brain MRI pretraining can support the prediction of anatomically meaningful embeddings. Nevertheless, the original NeuroCBIR embeddings remained the strongest representation across all metrics, indicating that the original preprocessing and embedding-generation pipeline still provides a measurable advantage.

5.2 Performance Across Dataset Subgroups

The subgroup analysis provides additional information about the robustness of the predicted embedding spaces.

Differences in retrieval performance were observed when the results were grouped according to scanner field strength. For the CNN, mS@3 increased from 0.828 for 1.5 T scans to 0.934 for 3.0 T scans, while PIP-Net increased from 0.840 to 0.962. In contrast, the original NeuroCBIR embeddings showed only a small difference, from 0.978 to 0.997. This suggests that the predicted embedding spaces were more sensitive to acquisition-related variability than the original NeuroCBIR representation. One possible explanation is that the NeuroCBIR preprocessing pipeline reduces scanner-dependent differences through bias-field correction, skull stripping, and spatial registration, whereas the proposed models must learn to handle these variations directly from minimally processed inputs. Nevertheless, the unequal number of queries in the two field-strength groups requires a cautious interpretation.

Differences were also observed across diagnostic groups, with both proposed models achieving their highest performance for cognitively normal subjects and substantially lower values for AD queries. However, the number of AD and MCI queries was considerably smaller than the number of CN queries. Moreover, diagnostic group may be associated with other factors, including the number of longitudinal acquisitions, scanner field strength, and disease-related anatomical variability. Therefore, these results do not necessarily indicate that the models are intrinsically less effective for Alzheimer's disease, but rather identify an aspect that should be investigated using larger and more balanced diagnostic subsets.

5.3 Generalization to External Datasets

The external evaluation was performed on MIRIAD, AIBL, and SLIM without additional fine-tuning. Overall, both models retained subject-specific information when applied outside OASIS-3, although performance differed considerably across datasets. PIP-Net achieved higher mean Success than the CNN on AIBL and SLIM, while both models obtained almost perfect results on MIRIAD. Some metrics are reported as N/A because the corresponding datasets did not contain enough longitudinal scans per subject to evaluate retrieval at larger values of k.

The high performance on MIRIAD should be interpreted with caution. This dataset contains several longitudinal acquisitions per subject. Scans from the same subject may therefore share highly stable anatomical and acquisition-related characteristics, making subject re-identification easier. The results show that the models preserved individual-specific information, but they do not necessarily imply equivalent performance in more heterogeneous clinical cohorts.

On AIBL, PIP-Net consistently outperformed the CNN, reaching an mS@1 of 0.827 compared with 0.696. This suggests that the prototype-based representation was more robust when applied

to a different cohort of older participants. Lower performance was observed on SLIM, particularly for the CNN. PIP-Net achieved an mS@1 of 0.604, whereas the CNN reached 0.381. SLIM represents the largest population shift because it contains healthy young adults, while the models were trained mainly on older subjects from OASIS-3. Differences in age distribution, anatomy, and acquisition characteristics may therefore have contributed to the lower retrieval performance.

These findings show that the learned representations retained useful subject-specific information across datasets with different populations and acquisition conditions.

5.4 Prototype-based Interpretability

The adaptation of PIP-Net to embedding regression preserved the possibility of inspecting spatial prototype activations, providing information about the local image patterns involved in constructing the predicted embedding. However, the interpretation of these prototypes differs from that of the original classification-based PIPNet3D. In the original model, a sparse layer with non-negative weights connects a limited number of prototypes directly to the predicted class. In the proposed model, the pooled prototype activations are instead combined through a linear layer to generate a 32-dimensional embedding. Consequently, individual prototypes cannot be interpreted as direct evidence for a diagnostic decision, and the resulting explanation is less compact than the scoring-sheet mechanism of the original architecture.

The qualitative analysis showed that 12 of the 20 prototypes with the largest contributions were mainly activated in regions containing identifiable brain anatomy. This indicates that the model learned several patterns related to cerebral structures rather than relying exclusively on global or non-anatomical image characteristics. Nevertheless, the remaining prototypes were activated near cortical boundaries, background regions, or the interface between the head and the surrounding image space. These activations suggest that the model may also exploit information related to head positioning, field of view, image boundaries, or acquisition conditions.

The absence of registration to a common anatomical space further limits the anatomical interpretation of the prototypes. The same brain structure may appear at different spatial coordinates across subjects because of differences in head orientation, brain size, and image geometry. Therefore, similar activation coordinates do not necessarily correspond to the same anatomical region, while different coordinates may still represent the same structure. The prototype visualizations should consequently be regarded as qualitative indications of the image patterns used by the model rather than as precise anatomical localizations.

Overall, the prototype-based architecture provides greater insight into the internal representation than the CNN baseline, whose embedding is produced from a global latent feature vector. However, the present analysis does not demonstrate that the model is fully interpretable or that the identified prototypes correspond to clinically meaningful biomarkers. A more complete assessment would require quantitative localization-consistency measures, registration to a reference atlas, and evaluation by neuroimaging experts. The results therefore support the use of PIP-Net as a partially interpretable embedding model, while also showing that further constraints are needed to obtain sparse, anatomically consistent, and clinically validated prototype explanations.

5.5 Methodological Limitations

Several methodological limitations should be considered when interpreting the results of this study. First, both proposed models were trained to reproduce the embedding space generated by NeuroCBIR rather than to learn anatomical similarity directly from the MRI volumes. Their performance therefore depends on the quality of the target representation. Any bias, loss of information, or acquisition-related pattern already encoded by NeuroCBIR may also be inherited by the predicted embeddings.

A further limitation arises from the use of minimally processed MRI scans. Avoiding skull stripping and spatial registration reduces the processing required for a new query, which was one of the main motivations of this work. At the same time, the input volumes retain variability related to head position, orientation, field of view, skull morphology, and background. The networks may consequently rely on information that is not exclusively related to brain anatomy. This aspect is especially relevant for the interpretation of PIP-Net prototypes, since activations located at similar coordinates do not necessarily correspond to the same anatomical structure across subjects when the scans are not aligned to a common reference space.

The adaptation of PIP-Net from classification to embedding regression also changes the nature of its explanations. In the original formulation, prototype activations are connected to diagnostic classes through a sparse and positive decision layer, making the contribution of each prototype relatively straightforward to interpret. Here, prototypes are instead combined to predict a multidimensional embedding. One prototype may influence several embedding components, while the final retrieval ranking depends on the interaction among multiple activations. The model remains more transparent than a conventional CNN because its internal evidence can be visualized, but the direct link between an individual prototype and the final retrieved cases is less explicit.

Sparsity is also only partially enforced in the current implementation. During training, the embedding prediction is computed from the complete set of pooled prototype activations. The activation threshold is applied only during inference, when values below 0.1 are set to zero. As a result, the model is optimized using a dense representation and evaluated using a partially sparse one. This difference may reduce the compactness of the learned explanation and does not guarantee that weakly activated prototypes become irrelevant during optimization.

The interpretation analysis itself was mainly qualitative. Only the prototypes with the largest contributions were examined in detail, and their anatomical relevance was assessed visually. No atlas-based localization metric or systematic evaluation by clinical experts was performed. The reported visualizations therefore show where the model responds, but they cannot by themselves demonstrate that the detected patterns correspond to clinically meaningful anatomical biomarkers. This is particularly important because some prototypes appear to capture boundaries, contrast variations, or other image characteristics rather than clearly identifiable brain structures.

Finally, retrieval performance was evaluated through subject re-identification. The results may therefore be influenced by the number of available scans per subject and by the interval between acquisitions, since scans obtained during the same session or at nearby time points may be easier to match than scans acquired several years apart.

5.6 Clinical Relevance and Future Work

The clinical relevance of the proposed approach lies in the possibility of generating retrieval embeddings directly from minimally processed MRI scans. Reducing the dependence on time-consuming preprocessing pipelines could make content-based retrieval more accessible in research and, in the longer term, in clinical settings. The system does not return only a numerical prediction, but a set of real examinations that can be compared with the query scan. This formulation may provide more tangible evidence than a conventional black-box output, although the retrieved cases still require interpretation by a qualified specialist.

Several aspects of the methodology could be refined in future work. A broader optimization of the architecture and training procedure may improve the agreement between predicted and target embeddings, particularly for the conventional CNN. This could include the evaluation of alternative backbones, embedding heads, loss functions, regularization strategies, and data augmentation methods. Training on multiple cohorts rather than on OASIS-3 alone could also reduce sensitivity to dataset-specific characteristics and improve robustness across scanner models, acquisition protocols, and patient populations. Another possible direction would be to

investigate an intermediate preprocessing strategy that preserves the practical advantages of minimally processed scans while reducing irrelevant variability caused by head orientation, field of view, and non-brain structures.

The prototype-based model presents several opportunities for further methodological refinement. Sparsity could be enforced during training by penalizing weak or unnecessary prototype activations. This would reduce the discrepancy between training, where all prototype activations contribute to the embedding, and inference, where activations below the threshold are set to zero. Reducing the number of weak or redundant prototypes may lead to more compact explanations without compromising retrieval performance. Their anatomical meaning should also be examined more systematically, for example through atlas-based localization, consistency measures across subjects, and evaluation by radiologists or neurologists. Such analyses would help distinguish prototypes representing meaningful brain patterns from those responding mainly to image boundaries, acquisition artefacts, or background information.

A further step is required to establish whether embedding similarity is clinically meaningful across different individuals. The current re-identification analysis mainly verifies that subject-specific anatomical information is preserved. Future studies should therefore evaluate whether patients retrieved as similar also share relevant morphological patterns, diagnosis, disease stage, cognitive characteristics, or clinical outcomes. Prospective evaluation and clinician-centred user studies would be particularly important before considering the system for decision-support applications.

Finally, the complete workflow could be implemented as an end-to-end research demonstrator. A user could upload the MRI scan of a new patient, generate its embedding, and retrieve the most similar cases from a reference database. The retrieved images could be displayed together with the available clinical information and, for PIP-Net, with the regions associated with the strongest prototype activations. Releasing the models, inference pipeline, and example interface in a public repository would facilitate reproducibility and allow other researchers to evaluate and extend the system.

6. Conclusion

The central question of this thesis was whether the retrieval representation learned by NeuroCBIR could be approximated without reproducing its full preprocessing pipeline. The experiments show that this is possible to a meaningful extent. Starting from minimally processed T1-weighted MRI scans, all three investigated models were able to generate embeddings that retained subject-specific information and supported image retrieval, although none fully reproduced the performance of the original NeuroCBIR representation.

The comparison between architectures also showed that the choice of model matters. The conventional CNN provided a useful baseline, but its performance varied considerably across the different validation-fold configurations. PIP-Net was markedly more stable and achieved higher overall retrieval performance, reaching an mS@1 of 0.865 ± 0.007 across the four configurations, compared with 0.613 ± 0.157 for the CNN. The pretrained FOMO2JOMO-based model reached results close to PIP-Net in the fold-1 comparison, suggesting that representations learned from large and heterogeneous brain MRI collections can transfer effectively to embedding regression. The original NeuroCBIR embeddings nevertheless remained the strongest representation, indicating that removing preprocessing simplifies the pipeline but does not come without a loss in retrieval accuracy.

The external datasets provided a different perspective on the same problem. Performance remained very high on MIRIAD, while the gap between PIP-Net and the CNN became more evident on AIBL and especially on SLIM. Rather than showing uniform generalization across cohorts, these results point to a dependence on population and acquisition characteristics. This is consistent with the subgroup analysis on OASIS-3, where the predicted embeddings were more sensitive than NeuroCBIR to differences in scanner field strength.

Adapting PIP-Net to embedding regression also made it possible to inspect part of the representation that would otherwise remain hidden inside a conventional CNN. Several of the most influential prototypes responded to regions containing identifiable brain anatomy, while others were associated with image boundaries, background, or head positioning. The prototype maps therefore provide useful evidence about what the model is using, but they should not yet be interpreted as anatomically precise or clinically validated explanations. The absence of registration and the dense contribution of multiple prototypes to the final embedding remain important limitations.

Taken together, the results support a simpler route toward MRI-based CBIR: NeuroCBIR-like representations can be predicted directly from minimally processed scans, and prototype-based

models offer a way to retain part of the transparency lost in standard deep regression. The next question is no longer only whether the same subject can be retrieved, but whether similarity between different subjects corresponds to clinically meaningful resemblance. Establishing that link, together with stronger prototype localization and evaluation across more heterogeneous cohorts, would be necessary before this type of system could move from a retrieval experiment toward a practical clinical support tool.

References

[1] World Health Organization. 2025. «Dementia». World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia

[2] Alzheimer's Disease International. «Dementia statistics». Alzheimer's Disease International. Retrieved May 27, 2026, from https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/

[3] National Institute on Aging. 2023. «Alzheimer's Disease Fact Sheet». National Institute on Aging. Retrieved May 27, 2026, from https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet

[4] Laboratory for Computational Neuroimaging. n.d. «FreeSurfer software suite». FreeSurfer. Retrieved May 27, 2026, from https://surfer.nmr.mgh.harvard.edu/

[5] ARAMIS Lab. n.d. «FreeSurfer processing». Clinica Documentation. Retrieved May 27, 2026, from https://aramislab.paris.inria.fr/clinica/docs/public/v0.6.0/Pipelines/T1_FreeSurfer/

[6] F. Nieto-del-Amor, J. Fu, J.-S. Muehlboeck, E. Westman, D. Ferreira and R. Moreno, "NeuroCBIR: A Fast and Accurate Image Retrieval System for Whole-Brain and Region-Specific MRI"

[7] Lane, C. A., J. Hardy, e J. M. Schott. 2018. «Alzheimer's Disease». *European Journal of Neurology* 25(1): 59–70. doi:10.1111/ene.13439

[8] Denver P, McClean PL. Distinguishing normal brain aging from the development of Alzheimer's disease: inflammation, insulin signaling and cognition. Neural Regen Res. 2018 Oct;13(10):1719-1730. doi: 10.4103/1673-5374.238608. PMID: 30136683; PMCID: PMC6128051.

[9] Dickerson BC, Sperling RA. Functional abnormalities of the medial temporal lobe memory system in mild cognitive impairment and Alzheimer's disease: insights from functional MRI studies. Neuropsychologia. 2008;46(6):1624-35. doi:

10.1016/j.neuropsychologia.2007.11.030. Epub 2007 Dec 8. PMID: 18206188; PMCID: PMC2760288.

[10] Kuperman, V., 2000. Magnetic resonance imaging: physical principles and applications. Elsevier

[11] Gossuin, Y., Hoca, A., Gillis, P., Vuong, Q.L., 2010. Physics of magnetic resonance imaging: from spin to pixel. Journal of Physics D: Applied Physics 43, 213001.

[12] Chandra A, Dervenoulas G, Politis M; Alzheimer's Disease Neuroimaging Initiative. Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. J Neurol. 2019 Jun;266(6):1293-1302. doi: 10.1007/s00415-018-9016-3. Epub 2018 Aug 17. PMID: 30120563; PMCID: PMC6517561.

[13] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," Journal of Medical Systems, vol. 42, no. 11, p. 226, Nov. 2018. [Online]. Available: http://link.springer.com/10.1007/s10916-018-1088-1

[14] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama,N. Teraya, A. Bolatkan, N. Shinkai, H. Machino, K. Kobayashi, K. Asada,M. Komatsu, S. Kaneko, M. Sugiyama, and R. Hamamoto, "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," Journal of Medical Systems, vol. 48, no. 1, p. 84, Sep. 2024.[Online]. Available: https://link.springer.com/10.1007/s10916-024-02105-8

[15] Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 29, 102-127

[16] Brusini I. Methods for the analysis and characterization of brain morphology from MRI images [doctoral thesis]. Stockholm: KTH Royal Institute of Technology; 2022. Available from: https://hdl.handle.net/10616/47924

[17] *Pan D, Zeng A, Yang B, Lai G, Hu B, Song X, Jiang T; Alzheimer's Disease Neuroimaging Initiative (ADNI). Deep Learning for Brain MRI Confirms Patterned Pathological Progression*

in Alzheimer's Disease. Adv Sci (Weinh). 2023 Feb;10(6):e2204717. doi: 10.1002/advs.202204717. Epub 2022 Dec 27. PMID: 36575159; PMCID: PMC9951348.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http: //arxiv.org/abs/1512.03385

[19] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Jan. 2018, arXiv:1608.06993 [cs]. [Online]. Available: http://arxiv.org/abs/1608.06993

[20] He S, Feng Y, Grant PE, Ou Y. Deep Relation Learning for Regression and Its Application to Brain Age Estimation. IEEE Trans Med Imaging. 2022 Sep;41(9):2304-2317. doi: 10.1109/TMI.2022.3161739. Epub 2022 Aug 31. PMID: 35320092; PMCID: PMC9782832.

[21] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60–88. Available from: https://doi.org/10.1016/j.media.2017.07.005

[22] Muehlboeck JS, Westman E, Simmons A. TheHiveDB image data management and analysis framework. Front Neuroinform. 2014 Jan 6;7:49. doi: 10.3389/fninf.2013.00049. PMID: 24432000; PMCID: PMC3880907.

[23] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "FastSurfer – A fast and accurate deep learning based neuroimaging pipeline," NeuroImage, vol. 219, p. 117012, 2020. doi:10.1016/j.neuroimage.2020.117012.

[24] Bashyam VM, Erus G, Doshi J, Habes M, Nasrallah I, Truelove-Hill M, Srinivasan D, Mamourian L, Pomponio R, Fan Y, Launer LJ, Masters CL, Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Satterthwaite TD, Wolf D, Gur RE, Gur RC, Morris J, Albert MS, Grabe HJ, Resnick S, Bryan RN, Wolk DA, Shou H, Davatzikos C. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain. 2020 Jul 1;143(7):2312-2324. doi: 10.1093/brain/awaa160. Erratum in:

Brain. 2021 Feb 12;144(1):e12. doi: 10.1093/brain/awaa328. PMID: 32591831; PMCID: PMC7364766.

[25] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel). 2020 Dec 25;23(1):18. doi: 10.3390/e23010018. PMID: 33375658; PMCID: PMC7824368.

[26] Viswan, V., Shaffi, N., Mahmud, M. *et al.* Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. *Cogn Comput* **16**, 1–44 (2024). https://doi.org/10.1007/s12559-023-10192-x

[27] Ahmed F, Naz NS, Khan S, Rehman AU, Ismael WM, Khan MA. Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. BMC Med Imaging. 2026 Jan 5;26(1):37. doi: 10.1186/s12880-025-02118-w. PMID: 41491470; PMCID: PMC12809972.

[28] M. Nauta, J. Schlötterer, M. van Keulen and C. Seifert, "PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 2744-2753, doi: 10.1109/CVPR52729.2023.00269

[29] L. A. De Santi, J. Schlötterer, M. Scheschenja, J. Wessendorf, M. Nauta, V. Positano and C. Seifert, "PIPNet3D: Interpretable Detection of Alzheimer in MRI Scans," *arXiv preprint arXiv:2403.18328*, 2024.

[30] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, M. E. Raichle, C. Cruchaga, and D. Marcus, "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease," Dec. 2019. [Online]. Available:http://medrxiv.org/lookup/doi/10.1101/2019.12.13.19014902

[31] A. Munk et al., "Towards Brain MRI Foundation Models for the Clinic: Findings from the FOMO25 Challenge," arXiv preprint arXiv:2604.11679, 2026, doi: 10.48550/arXiv.2604.11679

[32] I. B. Malone, D. Cash, G. R. Ridgway, D. G. MacManus, S. Ourselin, N. C. Fox, and J. M. Schott, “MIRIAD—Public release of a multiple time point Alzheimer’s MR imaging dataset,” *NeuroImage*, vol. 70, pp. 33–36, 2013, doi: 10.1016/j.neuroimage.2012.12.044.

[33] Liu, W., Wei, D., Chen, Q. *et al.* Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Sci Data* **4**, 170017 (2017). https://doi.org/10.1038/sdata.2017.17

[34] Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., ... Yastrubetskaya, O. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672-687. https://doi.org/10.1017/S1041610209009405