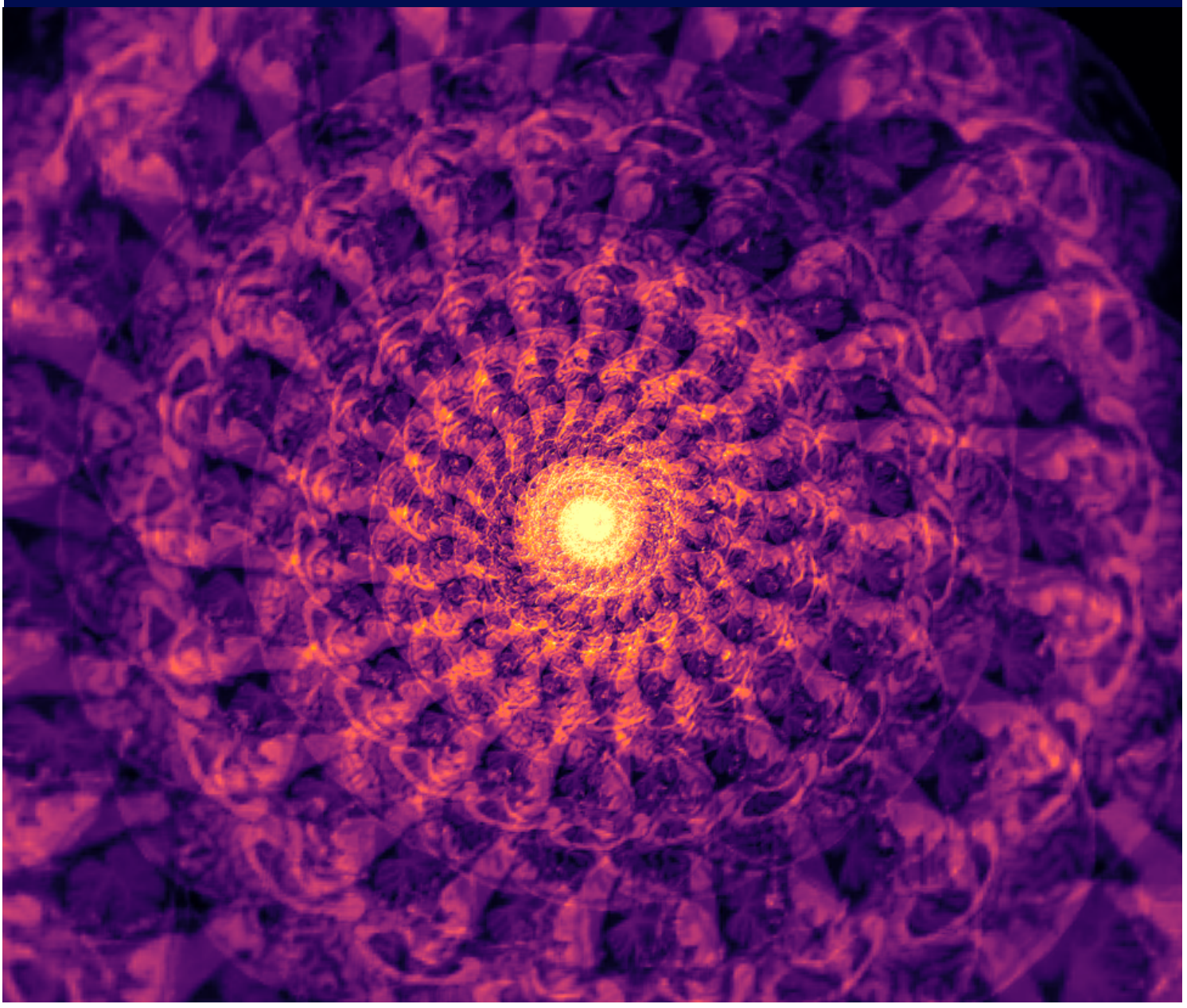
University of Stavanger

**JON HENRIK TJEMSLAND, KIRAN VÅGEN**
DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

# Unsupervised Deep Learning for Anomaly Detection of MRI Scans in Dementia

Bachelor's Thesis - Computer Science - May 2025



We, **Jon Henrik Tjemsland, Kiran Vågen**, declare that this thesis titled, “Unsupervised Deep Learning for Anomaly Detection of MRI Scans in Dementia” and the work presented in it are our own. We confirm that:

- This work was done wholly while in candidature for a bachelor’s degree at the University of Stavanger.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.

*“And then I said, oatmeal? Are you CRAZY?”*

– Pinke Pie

# Abstract

Alzheimer's disease (AD), the most common form of dementia, often manifests as structural changes in brain MRI scans. While machine learning models can detect these changes, they often rely on large amounts of labeled data. When the amount of data is limited, models tend to underperform outside the conditions in which they were trained, even if they appear accurate on similar test data.

This thesis investigates whether unsupervised learning could help overcome this limitation by analyzing brain changes without the need for labeled data, through the use of autoencoders to learn a latent representation of brain MRI scans and assess whether this latent space reflects meaningful biological and structural differences. While previous work has shown that such representations can be useful for classification tasks, less attention has been paid to the insights they can provide about the data itself.

Using a longitudinal brain MRI dataset of 594 individuals and an average of 3 scans per individual, we identify several statistically significant clusters. These clusters showed significant differences in the number of individuals with AD and mild cognitive impairment (MCI). Interestingly, we found that the latent space appeared to capture broader factors such as age and gender more strongly than the differences between AD and MCI. This suggests that the model prioritizes large-scale structural features over finer disease-specific details in order to minimize reconstruction error.

# Acknowledgements

We would like to thank our main and co-supervisors, Ketil Oppedal and Aksel Leknes, for their enthusiasm and valuable insights while writing this thesis. They helped guide us on the right path by frequently checking up on us and to give us a push when the hurdles seemed hard to overcome. Thank you.

We also thank each other, for our co-worker's input in writing this text. Thank you Kiran. Thank you Jon.

# Abbreviations

**AD Alzheimer's Disease**: A progressive neurodegenerative disorder and the most common cause of dementia.

**ADNI Alzheimer's Disease Neuro-imaging Initiative**: A large, multi-site longitudinal study that collects and shares clinical, imaging, genetic, and biochemical biomarkers for the study of AD.

**AE Autoencoder**: A type of neural network used to learn efficient codings of input data, often applied in anomaly detection or dimensionality reduction.

**ANN Artificial Neural Network**: A computational model inspired by biological neural networks, used in a wide range of machine learning tasks.

**CN Cognitively Normal**: Refers to individuals without cognitive impairment, often used as a control group in AD studies.

**CNN Convolutional Neural Network**: A type of neural network particularly effective for image data, commonly used in medical imaging tasks.

**MCI Mild Cognitive Impairment**: A clinical stage between normal aging and dementia, often considered a prodromal stage of AD.

**MRI Magnetic Resonance Imaging**: A non-invasive imaging technique used to visualize internal structures of the body, particularly the brain in AD research.

**SSIM Structural Similarity Index Measure**: A perceptual metric used to assess image quality by comparing structural information between images.

**VAE** **Variational Autoencoder**: A type of autoencoder that learns a probabilistic latent space, often used for generative modeling and unsupervised anomaly detection.

# Contents

**Abstract** iii

**Acknowledgements** iv

**Abbreviations** v

**1 Introduction** 1
- 1.1 Motivation . . . 1
- 1.2 Objectives and Contributions . . . 1
- 1.3 Outline . . . 3
- 1.4 Declaration on the Use of AI Tools and Ownership . . . 3

**2 Background** 7
- 2.1 Dementia . . . 7
- 2.2 Magnetic Resonance Imaging . . . 7
- 2.3 Anomaly Detection . . . 8
- 2.4 Convolutional Neural Networks . . . 9
- 2.5 Residual Networks . . . 9
- 2.6 Autoencoders . . . 9
- 2.7 Variational Autoencoders . . . 11
- 2.8 Autoencoders in MRI Anomaly Detection . . . 12
- 2.9 MRI Analysis Software . . . 13
- 2.10 Dimensionality Reduction Algorithms . . . 13
  - 2.10.1 Principal Component Analysis (PCA) . . . 14
  - 2.10.2 T-Distributed Stochastic Neighbor Embedding (t-SNE) . . . 14
  - 2.10.3 Uniform Manifold Approximation and Projection (UMAP) 14
- 2.11 Permutation Feature Importance . . . 14

2.12 Random Forest . . . 15
2.13 Clustering Algorithms . . . 15
2.14 Evaluation Metrics . . . 16
2.14.1 Reconstruction Metrics . . . 16
2.14.2 Clustering Metrics . . . 17
2.15 Related Work . . . 19

**3 Methodology** **20**
3.1 Dataset description . . . 20
3.2 Preprocessing . . . 22
3.2.1 MR Images . . . 22
3.2.2 Qualitative Factors . . . 24
3.3 Encoding . . . 24
3.3.1 Residual Networks Model . . . 24
3.3.2 Autoencoder Architecture . . . 26
3.3.3 Autoencoder Training . . . 26
3.3.4 Clustering Algorithms . . . 30
3.4 Existing Baselines . . . 32

**4 Results** **37**
4.1 Residual Networks Model . . . 37
4.2 Image Reconstructions . . . 38
4.3 Latent Space Encoding . . . 42
4.3.1 Dimensionality . . . 42
4.3.2 Information Loss in Downscaling . . . 44
4.4 Clustering . . . 46
4.5 Risk Factors . . . 54
4.6 Comparison to Existing Baselines . . . 54

**5 Discussion** **61**
5.1 Implications . . . 61
5.2 Limited Generalizability . . . 62
5.3 Limitations . . . 62
5.3.1 Imbalanced Dataset . . . 62
5.3.2 Training Time . . . 62
5.3.3 Longitudinal Data . . . 63

5.3.4 Outlier Removal . . . . . 63

**6 Conclusions** **64**
6.1 Experimental Results . . . . . 64
6.2 Future Work . . . . . 66
6.2.1 Blur in Reconstructed Images . . . . . 66
6.2.2 Data Augmentation . . . . . 66
6.2.3 Multimodal Models . . . . . 66

**A GitHub** **68**

**B Abbreviations of Features in the ADNI Dataset** **69**

# Chapter 1

# Introduction

This chapter introduces motivation, objectives and contributions as well as an outline and the statement of use of AI and ownership.

## 1.1 Motivation

Early and precise diagnosis of dementia has a large effect on the treatment received and patient quality of life, but is often hard to acquire due to limited resources in clinics. [7, 37] This deficit highlights the need for autonomous methods of detection. Reliance on expert clinicians creates a bottleneck due to limited resources and high costs.

Autoencoders represent a promising approach to analyze brain MRI scans without requiring large labeled datasets. [16] By leveraging these techniques, we can potentially detect rare diseases and structural abnormalities autonomously. [42] The ADNI dataset is a good candidate for unsupervised training for anomaly detection since it contains a lot of healthy subjects but relatively few cases of Alzheimer's disease. [1] By improving our understanding of how autoencoders interpret structural brain differences, we can refine these methods to complement clinical expertise and improving the quality of care.

## 1.2 Objectives and Contributions

This report investigates how autoencoders (AE) and variational autoencoders (VAE) can be leveraged to discover meaningful structural patterns and assist in early de-

tection of dementia.

This report seeks to give insight to the following questions:

- What is the impact of different latent space dimensionalities on reconstruction and clustering performance?
- Can reconstruction loss highlight specific individuals and or brain regions associated with neurodegenerative diseases, and how does this correlate with clinical knowledge?
- Does the learned latent space encode useful information for extrapolation and early detection of unlabeled diseases?
- How do architecture types (AE vs. VAE) and training loss functions (MSE vs. SSIM) influence reconstruction and clustering performance on the latent representations?

To address these objectives, we used a combination of dimensionality reduction techniques, clustering algorithms, and reconstruction based evaluation using autoencoder models trained on structural brain MRI scans.

Our contributions include a thorough exploration of the effect of various training parameters on the learned latent space, such as number of dimensions, loss function and model architecture as well as associating clustering behavior with biological, medical and physical features, improving interpretability on how structural changes occur in the brain.

All code and resources used in this thesis are publicly available on GitHub: https://github.com/jon-tj/DementiaMRI, except the MRI images, which are subject to restricted access through the ADNI repository and require an approved data use request.

## 1.3 Outline

- **Chapter 1: Introduction.** Discusses the motivation and contributions of this project.
- **Chapter 2: Background.** Explains several key concepts required for embarking on this research.
- **Chapter 3: Methodology.** Reports what was done to achieve the results presented in this report.
- **Chapter 4: Results.** An extensive list of results of the exploration.
- **Chapter 5: Discussion.** A brief discussion on implications of the findings and the limitations faced in the making of this report.
- **Chapter 6: Conclusion.** A summary of what was found, and a list of areas that require more work.

## 1.4 Declaration on the Use of AI Tools and Ownership

# Declaration on the Use of AI Tools for Assignment Submission

| | |
|---|---|
| Name | Jon Henrik Tjemsland and Kiran Vågen |
| Course Code or Name | DATBAC-1, Bachelor in Computer Science |
| Semester: | 6 |

---

At UiS, it is permitted to use AI-based tools properly unless otherwise stated in the course description or assignment.

In general, proper use of AI tools means that these tools are used as an aid. The response is prepared by the students themselves, and sources are cited in accordance with the regulations. Bachelor's and master's theses, as well as other submission assignments, are written and independent tasks. You are responsible for the final text submitted for assessment.

For more information on the proper use of AI tools: Examination | University of Stavanger

Fill out the form to indicate whether you have used AI tools in your assignment.

The purpose of describing this use is to show the reader how you have worked in your assignment. By doing this, you follow the requirements of transparency and integrity in an academic assignment, and it may be easier for the assessor to evaluate your own contribution. By being clear about how you proceeded, the assessor will get an explanation of how you have used language models in your assignment.

**Language editing:** Have you used generative artificial intelligence for proofreading/correction of parts or the entire text?

☒Yes
☐No

If yes, specify which tools you have used and briefly describe how you have used them in your assignment:

ChatGPT was used to improve language. Typically, the use of binding words and sentence structure to improve text flow and to generate latex tables and formulas that otherwise would take a lot of time to manually write in latex code. All improvements have been thoroughly evaluated and the better wording chosen to further restructure the text.

**Brainstorming:** Have you used generative artificial intelligence for ideas and suggestions for the structure of your assignment?

☒Yes
☐ No

If yes, specify which tools you have used and briefly describe how you have used them in your assignment:

Both ChatGPT and DeepSeek were used for educational help, text structure and wording. Brainstorming with the LLMs has been used to gain a broader overview of programming methods and solutions.

**Images and figures:** Have you generated one or more of the images/figures in the assignment using generative artificial intelligence?

☐ Yes
☒ No
☐ Not relevant

If yes, specify which tools you have used and briefly describe how you have used them in your assignment:

**Codes and algorithms:** Have you generated parts of the code/algorithms that 1) appear directly in the thesis/assignment or 2) have been used to produce results such as figures, tables, or numerical values, using generative artificial intelligence?

☒ Yes
☐ No
☐ Not relevant

If yes, specify which tools you have used and briefly describe how you have used them in your assignment:

ChatGPT, Deepseek and Claude have been used for code production, improvement and debugging. The AIs unfortunately can't produce large parts of code, but could in simple examples help do some of the repetitive time-consuming work such as adding labels to plots and writing a function that we could further modify to implement into the bigger codebase. This has let us spend more of our mental capacity on complicated and abstract tasks. These AIs are seen as a better alternative than google in finding information about programming.

**Other AI aids og tools:** Have you used other types of AI aids or tools in the creation of this assignment?

☐ Yes
☒ No

If yes, specify which tools you have used and briefly describe how you have used them in your assignment:

---

I confirm that I have read the information on the use of AI tools in connection with assignment writing and exams. Examination | University of Stavanger

**How to fill out the form if you have used generative AI/language models in your assignment:**

**Here are some examples of how a description might look:**

1. I have used the language model Copilot in this assignment for inspiration on the structure of my text. I used Copilot to suggest an outline, but I have only taken ideas from this outline, which I have reviewed and processed myself.
2. I have used the language model Copilot in this assignment for inspiration on the argumentation in my text. I have received input to develop arguments in the discussion section. I have not used Copilot to write entire paragraphs or sentences in the assignment. I have only taken ideas that I have reviewed and processed myself based on the syllabus material and other scientific sources I have searched for without using AI.
3. I have used Copilot to correct spelling and other grammatical errors and to ask for suggestions to simplify and clarify some short paragraphs. I have taken inspiration from these suggestions but written my own sentences instead of copying and pasting from the AI suggestions.

The labour was split as evenly as possible between the authors, with a few small exceptions:

- Subchapter 3.3.1 was largely the work of Kiran, implementing the model outlined in Phuong's repository [34], and documenting the effect of posterior collapse in 4.1.
- The effect of encoder-like models as a blur filter was documented primarily by Jon in 4.3.

# Chapter 2

# Background

This chapter provides an overview of the theoretical foundations relevant to the methods used in this thesis.

## 2.1 Dementia

Dementia is a disease characterized by a continuous decline in cognitive function, interfering with daily life. Alzheimer's Disease (AD) is the most prevalent type of dementia, accounting for 60–80% of all cases. [49] Structurally, AD is associated with characteristic neurodegenerative patterns, such as cortical thinning, hippocampal atrophy, and abnormally large ventricles. These structural changes reflect underlying pathological processes such as amyloid-beta plaque deposition and tau tangles. The development of these structural changes often begins years before symptoms are expressed, making early detection difficult using only non-invasive analysis techniques, as of today. [5]

## 2.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) provides a non-invasive method to visualize structural changes in the brain. Structural MRI captures high-resolution anatomical details that are vital in tracking disease progression and evaluating risks and the predictive power of biomarkers. One of the challenges in using MRI for machine learning is dataset heterogeneity. Variations in magnetic field strength (e.g., 1.5T vs. 3T), contrast, scanner manufacturers, and differences

in pre-processing steps taken in the individual analyses can significantly affect model performance and interpretation. Biological factors such as age, gender and race also play a large role in this variability and makes data sparsity a problem in machine learning on MRI data. [17]

## 2.3 Anomaly Detection

In many contexts, the definition of what is abnormal is unclear, whereas the definition of what is considered normal is well-defined. Anomaly detection algorithms are methods of using this definition to discover items that do not belong in a larger body of *normal* instances. In the context of brain MRIs, anomaly detection can flag structural patterns that do not match the general population, potentially indicating neurodegeneration or disease.

A common implementation of anomaly detection uses an artificial neural network (ANN) to predict either whether a sample belongs to the normal group or the anomaly group, or to predict what the corresponding *normal* item would look like and looking at the differences, through what is often called reconstruction loss. Items that do not belong to the normal class tend to get higher reconstruction loss and one can thus predict the probability of the item being an anomaly. The weakness of the ANN lies in the lack of interpretability it offers. Diagnosis is usually only the first step, being followed up by treatment to revert the anomaly. Borrowing from Walter et al [36], the low interpretability of these models become problematic because:

> "[...] even when an algorithm allows detection of patients and controls with high levels of accuracy, it can be difficult to establish which specific features of the data informed the categorization decision. Therefore, even in the presence of a successful algorithm, we may gain little or no mechanistic understanding of the disease under investigation. This limits the translational applicability of the findings, since the development of new treatments is normally informed by the underlying mechanisms." [36]

## 2.4 Convolutional Neural Networks

The Convolutional Neural Network (CNN) is a layer type of deep learning models based on the ANN, but in a convolutional (or rolling) manner. It can be viewed as a local ANN for each kernel (sliding window over the input) that outputs a new image, which is typically smaller unless padding is added. In image processing, CNNs are commonly applied to 2D and 3D data to make the input images spatially invariant. Data augmentation by rotating and scaling input images can further improve generalizability and transformational invariance but at the cost of longer training and inference times, and larger models. [39]

## 2.5 Residual Networks

Skip connections or residual connections are connections which allow the network to bypass one or more layers and pass the input directly to a deeper layer. This technique addresses the challenges of vanishing gradients and performance degradation when training very deep neural networks. The structure of a residual block can be expressed mathematically as:

$$y = x + f(x)$$

Where y is the output of a ResNet block, x is the input and f(x) represents the residual mapping which may come from a set of layers such as convolution, batch-normalization and ReLU activation. At first glance this may seem counterintuitive, as if we take the input x, blur it at f(x) and then add it back to x. But f(x) actually captures useful features or residuals such as edge highlighting, noise reduction, pattern detection, etc. This helps preserve information across layers and allows substantially deeper models to work. [31]

## 2.6 Autoencoders

Autoencoders (AE) are a classic model architecture introduced in the 1980s, used for various tasks such as anomaly detection, dimensionality reduction and other unsupervised learning tasks. AEs are architectures consisting of three parts; an encoder, bottleneck and a decoder. The encoder is responsible for compressing

the input into a latent representation by downscaling the input repeatedly until sufficiently small. Then, the bottleneck creates a dense representation by applying a final compression step, typically using a fully connected layer. During training, the AE passes the encoder outputs to the decoder, which reverts the encoding process and minimizes disparity between input and output of the AE.

The latent space the encoder produces, typically denoted as $Z$, often has a higher information density than the input. The encoding process can therefore be considered a type of feature extraction. In image processing, the down-sizing behaviour of the encoder is usually achieved by chaining CNN layers to extract spatial information into a more compressed representation.

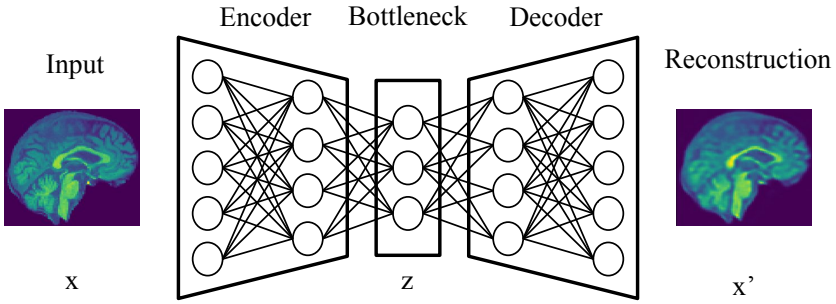



**Figure 2.1:** The autoencoder structure. The encoder compresses the input $X$ and passes it through a bottleneck layer, producing the latent space $Z$, which is reverted by the decoder into an approximation $X'$ of the input.

Expressing the encoding operation as $E$ and the decoding operation as $D$, the AE can be expressed as

$$X' = D(Z), \quad Z = E(X)$$

or more compactly as $X' = D(E(X))$. The loss function is normally defined as the mean squared error (MSE) of the reconstruction difference, also called the *reconstruction loss*, which can be expressed as

$$\mathcal{L}_{\text{recon}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - D(E(X_i))^2$$

Training the autoencoder entails finding the parameters of $E$ and $D$ that minimizes the loss function $\mathcal{L}_{\text{recon}}$. [28]

## 2.7 Variational Autoencoders

Variational autoencoders (VAE) extend on the AE architecture by learning a probabilistic model of the data. This is typically done by enforcing a distribution on the latent features, commonly the standard Gaussian distribution. This way, sampling the latent dimension using random normal samples generates new data points that exhibit some of the same characteristics of the dataset. Learning this probabilistic model is traditionally done by including a mean $\mu$ and variance term $\sigma^2$ in each of the latent dimensions, and combining the standard loss with a loss function enforcing the wanted distribution, using the Kullback-Leibler (KL) divergence which quantifies the difference between two probability distributions [23], expressed as

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\Big(q(Z|X) \;||\; p(Z)\Big) = \sum q(Z|X) \cdot log\Big(\frac{q(Z|X)}{p(Z)}\Big)$$

where $p(Z)$ represents the prior probability distribution, and $q(Z|X)$ is the approximate posterior, describing the distribution of $Z$ given an input $X$. The notation $||$ signifies a comparison between two distributions in terms of KL divergence.

The combined loss function to be minimized for a VAE is thus

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathcal{L}_{\text{KL}}$$

where $\beta \geq 0$ is the regularization strength hyperparameter. The special case $\beta = 0$ is a normal AE. In the original paper describing the VAE, $\beta = 1$ was fixed. A later study which introduced the beta-VAE where $\beta$ was a free hyperparameter showed the trade-off between regularization and learning. [19]

During training, the latent space must be sampled from the distribution to be handed to the decoder for reconstruction. This step introduces randomness to the latent variable, and is thus not differentiable. To ensure the gradients can be computed, the randomness introduced can be moved to an external variable $\epsilon$:

$$Z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

This step is known as the reparametrization trick and ensures differentiable gradients.

## 2.8 Autoencoders in MRI Anomaly Detection

AEs and VAEs can be used to find instances in a dataset that do not belong to a *normal* group by training a model to compress and reconstruct only the normal group. The mean reconstruction loss of the other groups will then statistically be higher than in the normal group if there is significant differences in input. In the case of MRI scans, the normal group is usually cognitively normal (CN), and other groups such as MCI and AD can then be reconstructed and will have higher loss than the CN group. This fact allows the model reconstruction loss to provide a useful metric of how different an instance is to the normal group. However, many challenges arise in applying standard anomaly detection techniques to these images.

High variability due to differences in gender, age, diet and genetics are often excluded from scientific papers due to scope limitations, sparking doubts about the generalizability of their findings. Unsupervised anomaly detection is vulnerable to high variability in inputs as it can hinder models from developing a coherent image of the relationships in the data. Distributional shifts due to non-pathological instances in the training data introduce significant drops in performance. This can be somewhat mediated by filtering out instances in the training set that is not similar to the other training instances, although this also loses variance. [10, 26]

Furthermore, the large volume of data contained in an MR image can rapidly lead to overfitting on in-sample data, further reducing generalizability. Analyzing out-of-sample images thus becomes a problem since the innumerable parameters of a trained model on such high dimensional input can often lead to the model "overestimating" its decisions, leading to a purely randomized output as the image is passed through the model. This can be remedied by regularization, but not removed, due to the nature of unsupervised learning.

The high associated cost of data acquisition also leads to few and often lower-quality data points (such as low resolution MR images). However, many of the subtle details characterizing the onset of AD and other diseases require a low-variability, high resolution dataset to consistently overcome the noise in the samples, particularly in unsupervised learning tasks, where the regions of interest can not explicitly be defined.

Since anomaly detection algorithms largely depend on a binary label for whether

an image is an anomaly, mislabeling by clinicians of MR images is another source of error. Images from ADNI are grouped by which clinical group the subjects belong to, leading to a binary label for whether an image belongs to the CN group or either of the MCI and AD groups. However, the structural changes associated with AD are continuous in nature, often leading to the model learning a blend of groups for either label.

## 2.9 MRI Analysis Software

The following software was used in the making of this report:

**FSL** is a library of analysis tools for FMRI, structural MRI and diffusion brain imaging data. It runs on macOS, Linux, and Windows via the Windows Subsystem for Linux. Most of the tools can be run both from the command line and as GUIs. The FSL tools used in this thesis are BET (Brain extraction tool), FAST (FMRIB's automated segmentation tool, used for bias field correction) and FLIRT (FMRIB's linear image registration tool, used for spatial normalization). [32]

**Synthstrip** is an advanced skull stripping tool designed to extract brain voxels across diverse image types, and the tool used in this project. Through the use of a deep learning approach, it synthesizes training images directly from segmentation maps, which has produced a highly generalizable model that remains invariant to acquisition parameters. [3]

## 2.10 Dimensionality Reduction Algorithms

As autoencoders produce a latent space representation in n-dimensions of each brain, whereas human vision is limited to at most 3-dimensional representations, it is necessary to reduce the number of dimensions to give an intuitive visualization of the encoding. Since this article is intended for 2-dimensional viewing, we need algorithms that can project the encodings to 2-dimensional space. This project applies PCA, t-SNE and UMAP for projecting latent spaces, which are further explained below.

### 2.10.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a linear dimensionality reduction method that projects data onto orthogonal axes of maximal variance. It preserves global structure at the cost of local relationships, it is fast to compute and is suitable for initial data exploration. As a linear dimensionality reduction algorithm, it cannot capture non-linear relationships. The success of the decomposition therefore depends on the n-dimensional input to be properly pre-processed to make it more linear. [21]

### 2.10.2 T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-Distributed stochastic neighbor embedding (t-SNE) is a nonlinear technique focusing on local structure. It preserves neighborhood relationships through probability distributions and is excellent for visualization at the cost of being computationally intensive. It offers a perplexity parameter that controls the balance of local and global patterns. [45]

### 2.10.3 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a nonlinear technique, grounded in Riemannian geometry and algebraic topology. UMAP competes with t-SNE in visualization quality, offering faster runtime performance, and improved scalability. Unlike t-SNE, UMAP focuses more on retaining more of the global structure, making it well suited for exploratory data analysis. [27]

## 2.11 Permutation Feature Importance

In many cases it is of interest to analyze the importance of features to a regression model. To do this it is possible to shuffle, or *permute*, a feature or a subset of features and examine how the permutation affects the performance of the regressor using some metric (typically MSE 2.14, $R^2$ or accuracy score); if the performance metric deteriorates after permutation, the features that were permuted are assumed to convey important information to the regressor model.

In certain cases, this drop in performance may also be due to an over-reliance of the model on a small subset of features, however, by aggregating many runs by

many regressor models, it is possible to gauge the expected drop in performance and, by extension, the *importance* of the feature to a regressor model.

A key weakness of permutation feature importance is correlated features. When many features convey the same or redundant information, the importance of each feature by itself is drastically reduced, since the model can in effect reconstruct the feature using the remaining unpermutated features. To mitigate this, correlated features can be removed or grouped together in the permutation process, however, knowing which features to group together is often not straightforward, potentially leading to confirmation bias. [6]

## 2.12 Random Forest

Random forests are a non-linear learning method used for classification and regression tasks. It works by constructing a large number of decision trees during training and outputting the average prediction for regression or the majority vote for classification across the ensemble during inference.

One of the key advantages of random forests is their ability to capture complex feature interactions in a manner that is easy to interpret. However, random forests may sometimes struggle with very high-dimensional data or when the data set contains many irrelevant or redundant features. Another weakness of random forests is that they rapidly become computationally expensive as the number of trees and dataset size grow, and are therefore most often used for analysis or low-dimensional tasks. [8]

## 2.13 Clustering Algorithms

**K-means** is a centroid-based linear clustering algorithm which assumes that clusters are spherical and similarly sized. It works well with compact and linearly separable data and is widely used due to its simplicity, scalability and efficiency. However, it can struggle with clusters of varying shapes, densities, sizes and is sensitive to initial centroid placement. [30]

**Spectral Clustering** is a graph-based non linear clustering algorithm which captures complex structures through the use of graph theory to exploit data connectivity. The limitations are that it is computationally expensive and requires

tuning affinity parameters. [14]

## 2.14 Evaluation Metrics

To evaluate the performance of our models, we used both reconstruction-based and clustering-based metrics. These metrics help quantify how well the autoencoders preserved input information and how effectively the learned latent spaces separated distinct data clusters.

### 2.14.1 Reconstruction Metrics

To assess the quality of image reconstruction, we employed two standard metrics: Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM).

**Mean Squared Error (MSE):** MSE measures the average squared difference between corresponding pixels in the original and reconstructed images. It is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where $y_i$ and $\hat{y}_i$ are the original and reconstructed pixel intensities, and $N$ is the total number of pixels. [15]

**Structural Similarity Index Measure (SSIM):** SSIM evaluates the perceived quality of the reconstructed images by considering luminance, contrast, and structural information. It is calculated as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- $\mu_x, \mu_y$ are the means of images $x$ and $y$,
- $\sigma_x^2, \sigma_y^2$ are the variances,
- $\sigma_{xy}$ is the covariance between $x$ and $y$,

- $C_1, C_2$ are small constants to stabilize the division when the denominator is close to zero.

SSIM values range from $-1$ to $1$, where $1$ indicates perfect similarity. Note that this metric only works for 2D images and 3D MRI scans thus have to be cut up such that each slice can be measured separately and then averaged. [29]

### 2.14.2 Clustering Metrics

To evaluate how well the latent representations grouped similar data points, we used the following clustering metrics:

**Silhouette Score:** Silhouette score measures how similar a point is to its own cluster compared to other clusters. The silhouette score $s$ for a single sample is defined as:

$$s = \frac{b - a}{\max(a, b)}$$

where:

- $a$ is the mean intra-cluster distance (how close the point is to other points in the same cluster),
- $b$ is the mean nearest-cluster distance (how close the point is to points in the nearest neighboring cluster).

The score ranges from $-1$ (poor clustering) to $1$ (well-clustered). A score of at least 0.25 is normally required to assume clustering is real and not just overfitting. [22]

**Inertia (Within-Cluster Sum of Squares):** Inertia quantifies the compactness of clusters by summing the squared distances of each point to its assigned cluster centroid:

$$\text{Inertia} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where $C_i$ is the $i$-th cluster and $\mu_i$ is its centroid. Lower inertia indicates tighter and more compact clusters. [11]

**Adjusted Rand Index (ARI):** The Rand Index (RI) is a measure of pairwise accuracy between two clusterings, defined as:

$$\text{RI} = \frac{\text{TP + TN}}{\text{TP+TN+FP+FN}}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. Unlike the raw Rand Index, ARI adjusts for chance groupings, making it more reliable:

$$\text{ARI} = \frac{\text{RI} - \text{Expected}_{\text{RI}}}{\max(\text{RI}) - \text{Expected}_{\text{RI}}}$$

where RI is the Rand Index, and $\text{Expected}_{RI}$ is its expected value under a random model. And max(RI) is by default 1.

ARI ranges from $-1$ to $1$, where:

- $1$ indicates perfect agreement between predicted and true labels,
- $0$ implies random labeling (expected value),
- Negative values suggest worse-than-random clustering.

[44]

**Clustering Accuracy** Clustering accuracy measures how well the predicted labels match the true class labels. It is defined as the ratio of correctly classified samples out of the total number of samples.

$$\text{Accuracy} = \frac{\text{Corrrectly classified labels}}{\text{Total labels}}$$

A higher Accuracy or ARI indicates that the latent space has been structured meaningfully, successfully separating distinct clinical groups without direct supervision. This can mean that the learned latent features capture relevant pathological patterns, making the model potentially effective for downstream tasks like anomaly detection.

## 2.15 Related Work

Previous work in brain MRI analysis has focused on supervised classification, often using hand-crafted features such as Hu moment invariants in SVMs to detect conditions such as alcoholism [25, 33], or CNNs in data augmentation for performance improvement [38]. Recent unsupervised methods, including autoencoders [35], GANs [18], and contrast-predicting models [4], detect structural anomalies through image reconstruction or transformation. These approaches are primarily aimed at differentiating healthy and diseased brains and do not address disease heterogeneity or subtype discovery.

Our approach uses a deep autoencoder not only to detect anomalies, but to learn the latent representation of brain structure on MRI scans, which potentially allows clustering of patients into subtypes. In contrast to previous works, such as Pinaya et al. [35], who used reconstruction error to identify regional anomalies, we cluster subjects in the latent space to uncover underlying structure in the dementia population, independent of diagnostic labels. This allows us to go beyond binary or multiclass classification and address the known clinical heterogeneity of Alzheimer's disease and related diseases.

By applying this method to the ADNI dataset, we leverage its rich multimodal data for potential validation and interpretation of the identified clusters. Unlike slice-based GANs [18] or contrast prediction models [4], our latent representations of the whole image allow for a global and integrative view of brain structure. Our method also builds on ideas from spatial autocoding in Parkinson's research [46], but our work is focused on subtyping in dementia. This positions our model as a tool to identify important subgroups that may be associated with diagnosis, biomarkers, or treatment response.

# Chapter 3

# Methodology

This chapter describes the materials, methods and workflow used in this project, including preprocessing, the proposed approach and implementation details.

## 3.1 Dataset description

This project uses data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), is a longitudinal, multi-center, observational study. The overall goal of ADNI is to validate biomarkers for Alzheimer's disease (AD) clinical trials. [1]

The dataset comprises 2,000 MRI scans from 594 individual subjects, along with comprehensive background information, including biological, physical, and medical characteristics. Table 3.1 provides a breakdown of participants by gender and diagnostic group. Figure 3.1 shows the correlation between some major descriptors of participants, showing a somewhat unbalanced dataset. Note that MMSE (the feature called MMSE_INTERP) is used in setting the AD diagnosis and is therefore strongly (negatively) correlated, as expected. However, we see that age (Age) and years in education (PTEDUCAT) are also very strong predictors of AD. This introduces biases in models trained on the dataset, which may or may not be the wanted effect.

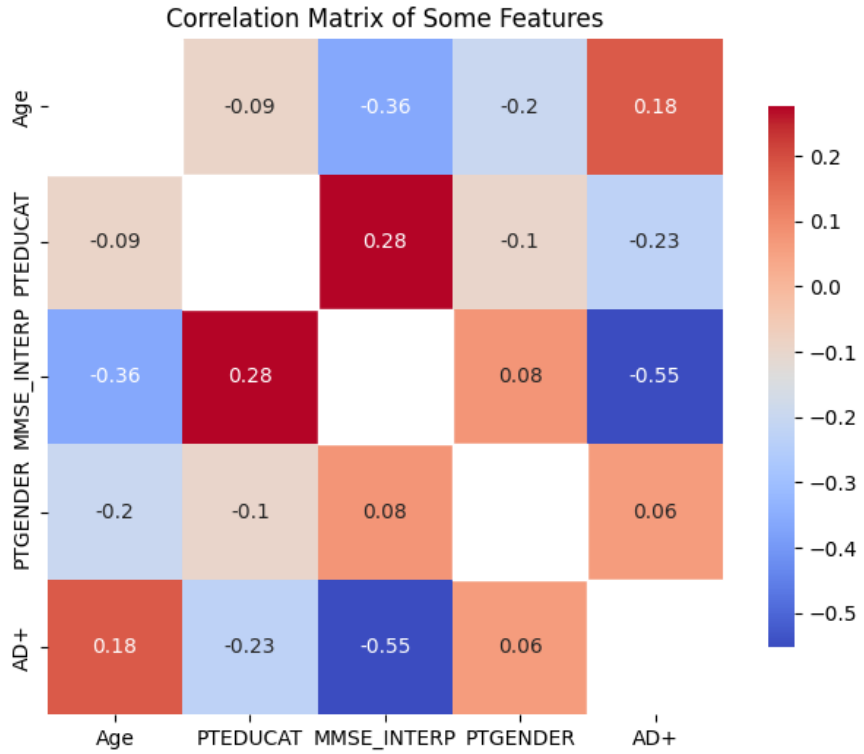



**Figure 3.1:** Feature correlation matrix. MMSE score and age are the strongest predictors of being Alzheimer's disease positive (AD+).

| Group | PTGENDER | Image count | Subject count |
|---|---|---|---|
| AD | Female | 132 | 16 |
| | Male | 81 | 8 |
| CN | Female | 780 | 303 |
| | Male | 546 | 210 |
| MCI | Female | 145 | 20 |
| | Male | 316 | 37 |
| | **Total** | **2000** | **594** |

**Table 3.1:** Number of participants in the dataset grouped by diagnostic group and gender. The AD group contains far fewer participants than the CN group, and is heavily skewed to females. The CN group also contains more females than males but not as much as AD.

## 3.2 Preprocessing

### 3.2.1 MR Images

The images downloaded from ADNI have already undergone some preprocessing steps. This is to ensure consistency and accuracy across different websites over time which also helps encourage collaboration and comparisons of research and models. The already applied preprocessing steps include [2]:

- **Gradwarp correction:** Revert back distortions due to gradient non-linearity which varies with each specific model.
- **B1 non-uniformity correction:** Correcting image intensity non-uniformity that results from RF transmission.
- **N3 non-uniformity correction:** A histogram peak sharpening algorithm.

Removing information that lies outside the scope of the analysis before training the model is vital to produce a clean latent space that captures important features. A natural first step is the separation of the brain from the skull and the neck, for which we used Synthstrip. Further variability can be removed by undoing the rotation and positional offsets of each subject in the images. Image registration is the process of aligning images to a template to remove this variability, thus further increasing the information density of the latent space we will produce. For this, we used the FLIRT module from FSL. [24]

After the images were stripped and registered, we cut the image to the minimum bounding box containing the brain, to minimize the impact of a noisy background and reduce computation time.

An overview of the final pre-processing pipeline, along with some attempts that provided unsatisfactory results, can be seen below. By pre-processing the images before training, we reduce the information that must be retained in the latent spaces, hopefully increasing the ratio of useful information for our analysis.

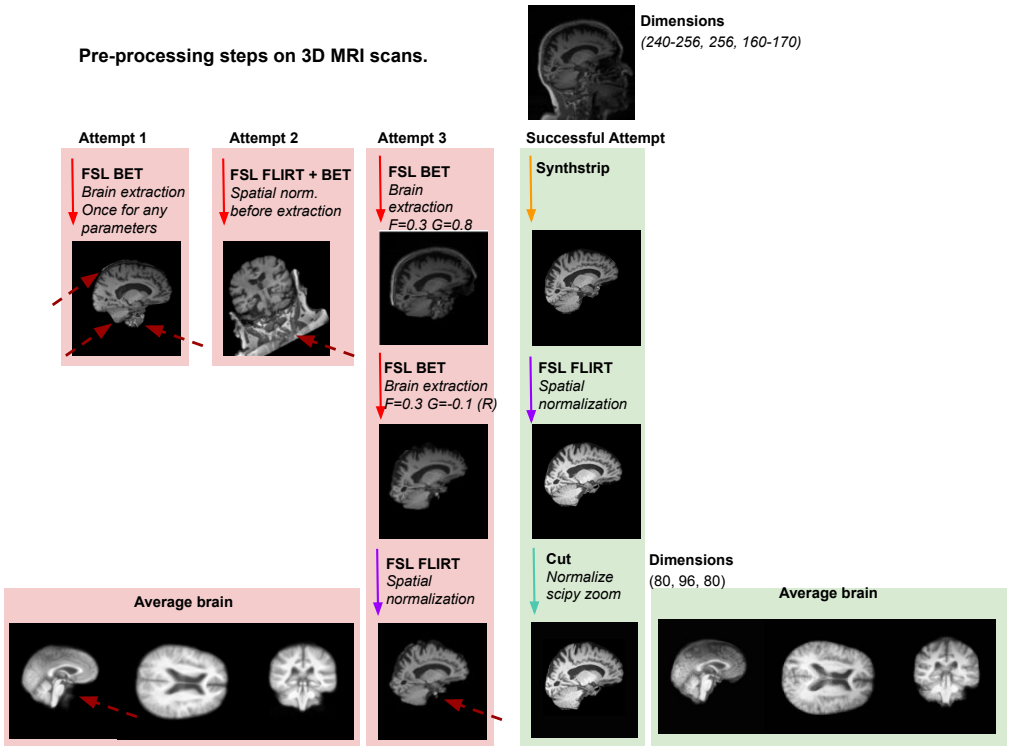



**Figure 3.2:** Iterations of the pre-processing pipeline. Four attempts were made to extract the brains from the skulls using a combination of FSL and Synthstrip. The final pre-processing pipeline used Synthstrip for extraction and FLIRT for registration, before being cut to contain only the brain and down-scaled for faster processing.

Although FLIRT works perfectly for spatial normalization, BET proved to be a difficulty because one either had to let some of the skull be left or one had to cut small fractions of the brain. This was a big problem as it also seemed to slightly mess up spatial normalization. Another problem was that if the tilt of the head was too high, no combination of hyperparameters seemed to work and running spatial normalization before performing skull stripping actually made it worse because some of the images contained even the shoulders of the patient. After trying to adjust the gradient intensity (G) and fractional intensity (F) to each image, a third attempt was made by simply performing the skull stripping in two steps. First removing the neck, then removing the skull. This gave good results, however, manual inspection revealed the original images had different rotations, tilts and sizes, necessitating a better solution. We found Synthstrip to work no matter the rotation, tilt and size of the original images, and it could cut out the brains

perfectly without having to adjust any input parameters. Combining Synthstrip for skull-stripping with FLIRT for spatial normalization gave the best results.

### 3.2.2 Qualitative Factors

Further processing was necessary on the dataset regarding entries in the biological (age, gender, etc.), physical (height, weight, etc.) and medicinal columns to create a useful background analysis for what could feasibly cause the differences in disease representation. In the raw dataset, missing values are frequent due to difficulties in data collection, and appear in various forms: inconclusive test results were commonly coded as -4 for numerical features or left blank otherwise; untested subjects were also recorded with blanks. Given the large dataset, it was feasible to impute some missing values based on available data per subject. However, certain features, like left-handedness, had too much uncertainty for reliable imputation. In such cases, we chose to drop records with missing values to maintain confidence in the dataset.

## 3.3 Encoding

### 3.3.1 Residual Networks Model

A model architecture based on residual network (ResNet) blocks was found in a public repository by Dao Duy Phuong [34] and implemented for a variational autoencoder. Both the convolutional and ResNet block consisted of Conv3D layer, a BatchNorm3D layer, and a ReLU activation. The key difference is that the ResNet block includes a skip connection, feeding both the output of the activation layer and the original input forward to the next block. The upsampling block consisted of Conv3D and UpSample layers. These components, together with MaxPool3D, were used to construct the encoder and decoder. The complete architecture is shown in Figure 3.3.

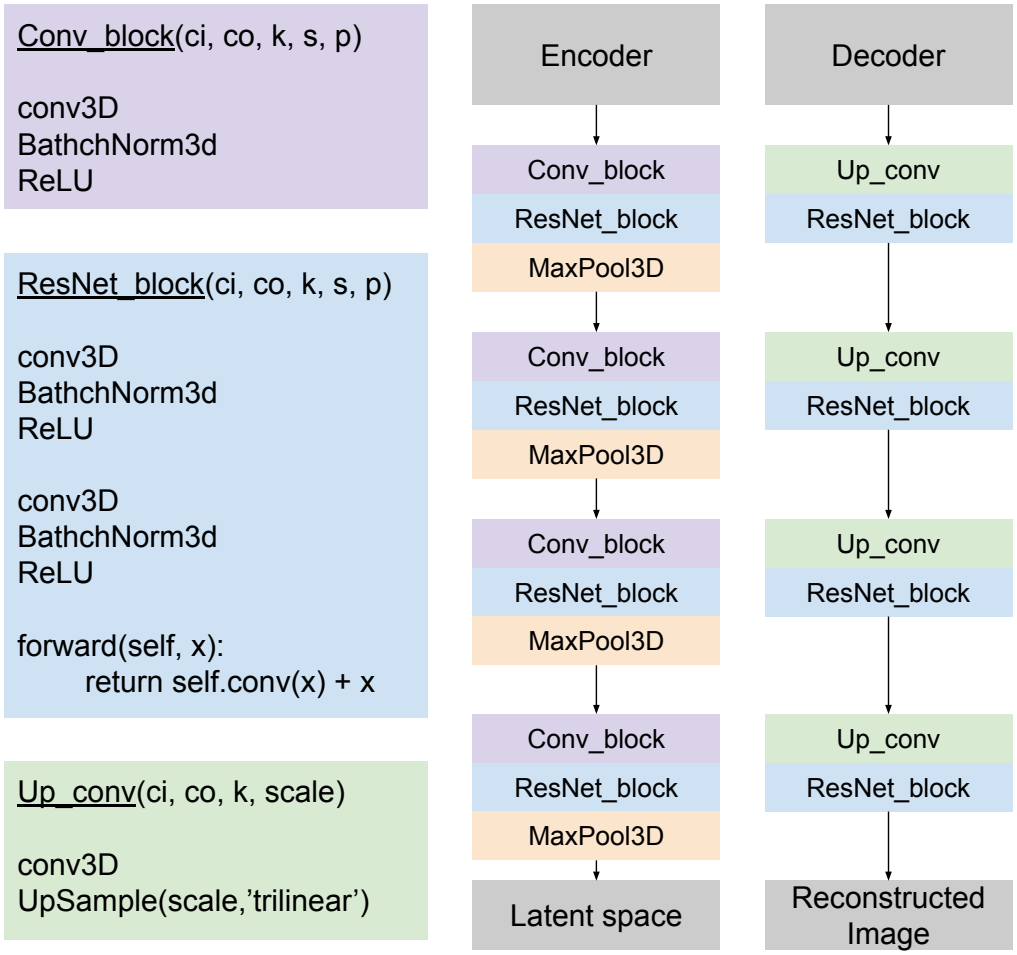



**Figure 3.3:** First encoder architecture utilizing residual networks.

Most of the parameters in the code were hardcoded, which limited its flexibility and made it difficult to experiment with. To address this, modifications were made to allow input of different shapes and sizes. For efficency during testing, inputs were also downscaled by factors between 2x and 4x. During training, several hyperparameters were explored, including input size, batch size, learning rate, weight decay, and latent space dimensionality. Additionally, the $\beta$-VAE formulation was implemented by introducing a $\beta$ multiplier to the KL-divergence term in the loss function. This was motivated by the need to reduce the strength of the KL term, a common strategy to address issues related to posterior collapse.

Due to persistent issues with maintainability and extensibility, largely caused

by the rigid hardcoding of parameters, this model was eventually set aside in favor of developing our own, more adaptable architectures.

### 3.3.2 Autoencoder Architecture

Figure 3.4 gives an overview of the autoencoder architecture used in this project. Optionally, various regularizers, including L1 and L2 regularization, could be applied to both the convolutional layers and the dense layers, keeping in line with the exploratory nature of this project.

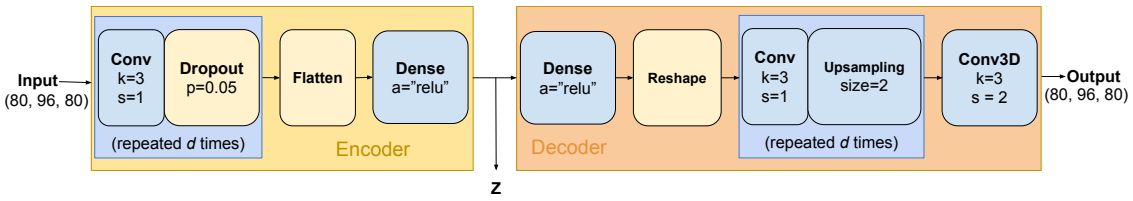



**Figure 3.4:** Using repeated convolutional layers to downscale the image while maintaining the spatial information stored in it by expanding the number of channels as we go, we can produce a spatially unaware representation that then gets flattened and fed into a dense layer to produce the final compact encoding. During downscaling, a Dropout layer is utilized to make the model resilient to overfitting.

A Flatten layer is used between the 3D output of the convolutional layers and the dense network to explicitly remove spatial information, supporting the exploratory nature of the work and allowing the model to define its latent space without being constrained by the spatial structure of the input data. Spatial constraints can limit the model's ability to discover abstract patterns, which is undesirable in this context. By flattening the data and learning features in a fully connected (dense) layer, the model can discover global patterns and learns the pattern of the inputs with less compute and fewer samples than a CNN, which can lead to better compression and more efficient feature extraction compared to an architecture that preserves spatial relationships in the latent space. [12]

### 3.3.3 Autoencoder Training

During the training phase of an AE that will be used for anomaly detection, the model only trains on the normal group. The MRI scans of the CN group were split into train and test sets at a 15% test split. Feeding the train images to the

autoencoder architecture outlined above, we produced latent representations of each image. To verify that the autoencoder successfully can reconstruct MR images of healthy (CN) subjects, thus providing a useful latent space representation, we can look at the reconstruction loss of the test set, and the residuals across each group.

Figure 3.5 displays the reconstructions and reconstruction loss for two individual brains sliced in the sagittal plane, as well as the mean input and reconstructions of one model trained on this data. The reconstructions are fairly similar to each other and to the mean reconstruction, indicating a lack in captured variance by the autoencoder. The reconstruction loss highlights the areas of most variance loss, in general highlighting the ventricular areas and the corpus callosum, indicating large individual variance in these areas. After 60 epochs, the mean squared error (MSE) on the training set ($0.0036 \pm 0.00031$) and validation set ($0.0040 \pm 0.00022$) metrics exhibited low variance, indicating consistent model performance across experiments, independent of hyperparameter choices, and minimal overfitting. Many hyperparameter combinations were tested to see the effect they have on clustering.

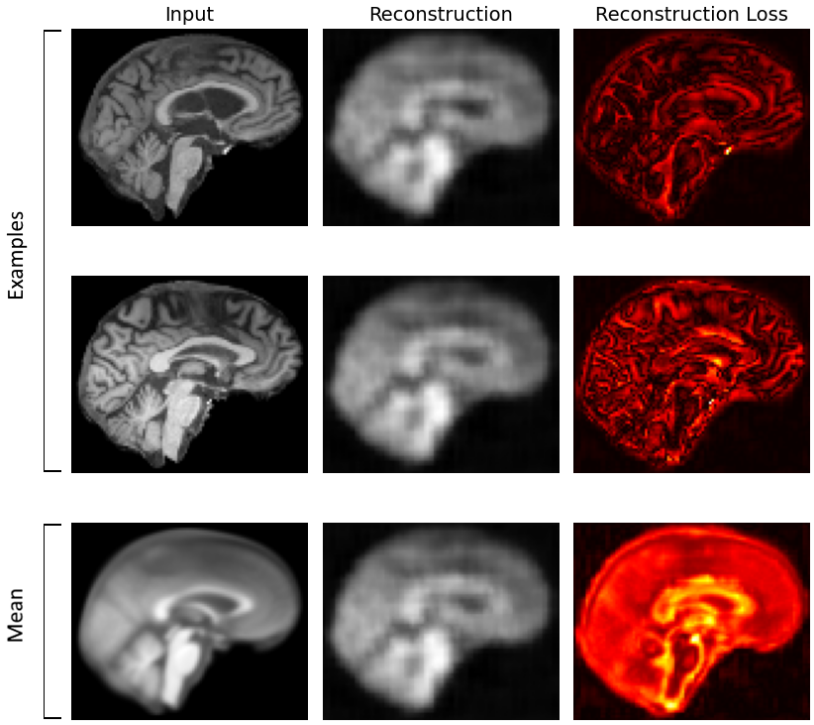



**Figure 3.5:** Reconstructions and reconstruction loss for autoencoder on the CN group of the dataset. Two individual images have been reconstructed as examples, and the mean of all inputs and reconstructions is given in the bottom row. This particular model had 256 latent dimensions and 4.93M trainable parameters.

The sharpness of the output image from the decoder correlates to how familiar the convolutional layers are about this specific pattern. This means that the autoencoder functions similar to a low-pass filter by blurring out regions it is not certain about. The function of the autoencoder as a low-pass filter can be proven by looking at the frequency domain of the input vs the output images. Figure 3.6 displays how the higher frequencies of an image passed through such a model become attenuated, similar to how applying a pure low-pass filter would do.

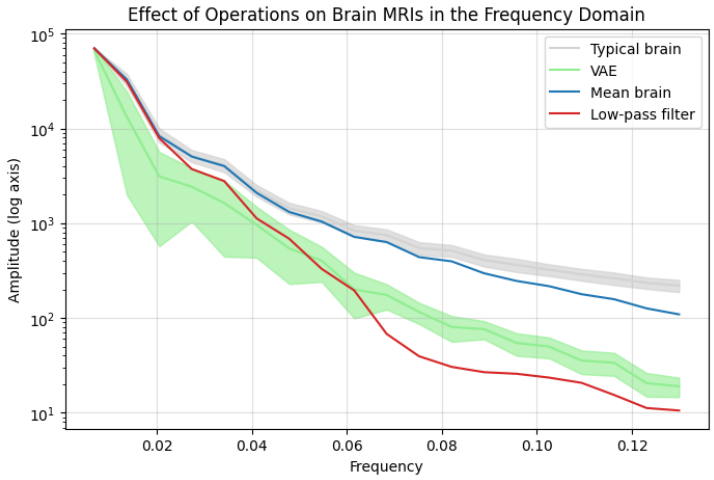



**Figure 3.6:** Applying an autoencoder like model to reconstruct an input image has a similar effect to applying a low-pass filter such as gaussian blur, or taking the longitudinal mean of all the images in the frequency domain of the image; attenuating higher frequencies. **Typical and mean brain label meanings:** Typical brain = AE on CN, Mean brain = AE on CN, MCI and AD.

This may be troublesome since some of the residuals in the AD group are known to be high-frequency, such as the smoothing of brain folds, which are high-frequency details. This blurring effect can be combated by penalizing finer details more than their large-scale counterparts, i.e. by increasing their amplitude before training the model, however doing so results in more noisy reconstructions, lowering the threshold between the groups. The models trained for this project already have small margins of error that cannot be eroded by noise. Due to scope and time constraints of the work, further work in exploring the importance of these fine details were left out. [1]

The reconstruction error on the normalized images have been summarized in table 3.2 by group. Note that the CN group has a lower error than the MCI and AD groups, giving some margin to distinguish the groups using only reconstruction loss.

[1]The fact that the decoder is unable to reconstruct these fine details does not necessarily mean the encoder is incapable of encoding them, it could also be the case that the decoder cannot reconstruct the complex pattern. However, since we do not observe these details being reconstructed, we also have no proof that the encoder does encode them.

**Table 3.2:** MSE loss and relative loss over all images.

| Quantile | CN (MSE) | CN/CN | MCI/CN | AD/CN |
|---|---|---|---|---|
| 0.5 | 0.00039238 | +0% | +2.3% | +9.1% |
| 0.95 | 0.02020991 | +0% | +28.6% | +33.0% |

Throughout the making of this report, many models with slightly different hyperparameters were trained to study the effect of tweaking parameters. Among other things, this included the dimensionality (cardinality) of the latent space, ranging from 3 to 3000 variables, the number of parameters of the convolutional layers, ranging from about 10k to 10M parameters, and the amount of regularization used for both L1 and L2 loss, and for KL-divergence (in the case of variational autoencoders). Since training deep learning models is an inherently randomized task, we must perform multiple runs of the training process to get an overview of the minimized loss on the model's loss manifold. Due to compute constraints, repeat runs were focused on the "baseline" model, which could then be compared with only a few runs of a modified model.

### 3.3.4 Clustering Algorithms

Having confirmed the autoencoder provides useful information in the latent space, a clustering algorithm could be made to identify subgroups in the population that belong to either a) almost entirely a single clinical diagnosis, thus potentially providing insight into what makes this group different from the others, or b) to a group of subjects that are very different from the others in the study, which can potentially be explained by other factors in the ADNI dataset to further study how genetical, physical and medicinal factors can increase the risk of AD or the risk of misdiagnosis, depending on the diversity in the subgroup. Before a choice of clustering algorithm to use for the task could be made, it was important to know the distribution of points in the latent spaces produced by the autoencoders. To gain insight into their distributions, t-SNE was used to project the n-dimensional latent spaces into 2D for plotting, which can be seen in figure 3.7.

From these scatter plots we can see that a simple K-means clustering algorithm may provide unsatisfactory clustering results, since the scatter plots clearly display lines of data points lying close to each other, indicating local directionality is important, whereas K-means is non-directional.

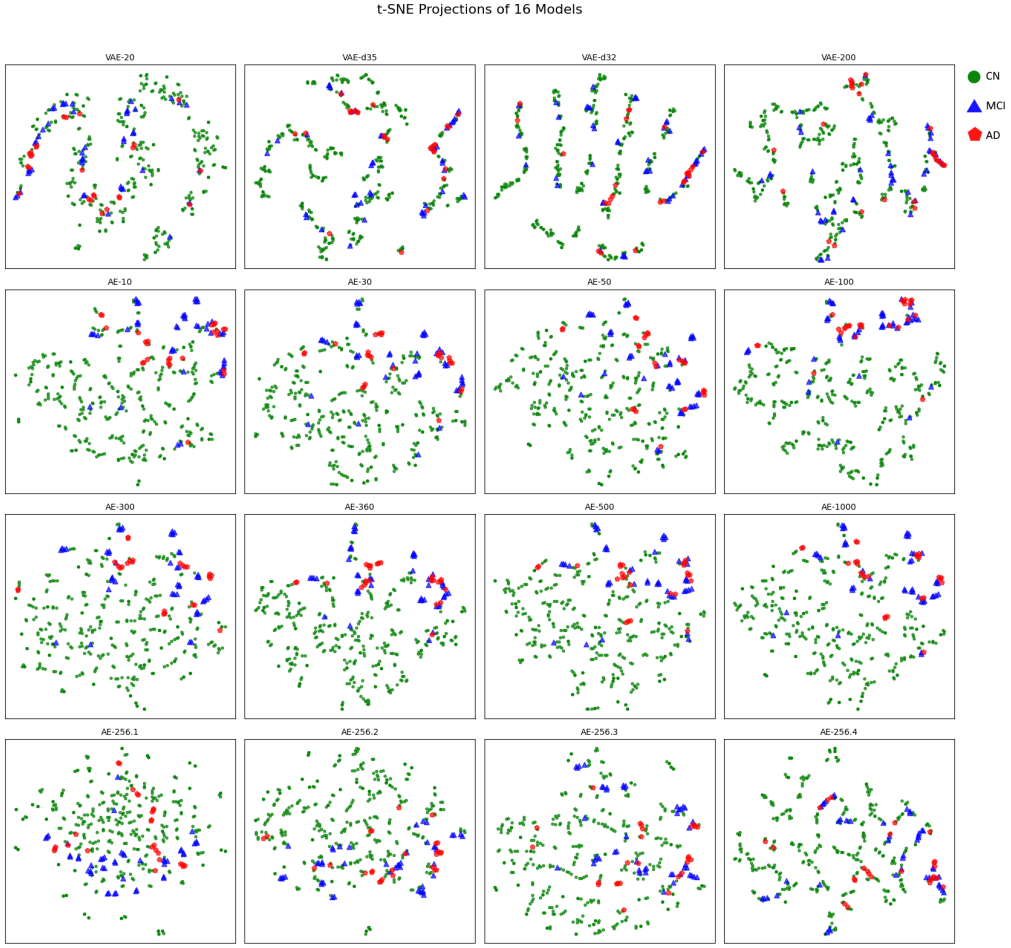



**Figure 3.7:** Scatter plot of latent dimensions projected into 2-d. The scatter plots display lines of data points lying close to each other, indicating a K-means approach may fail to capture the variability. **Participant groups:** Green circle — CN. Blue triangle — MCI. Red pentagon — AD.

To increase reliability of our analysis, we used multiple clustering algorithms including K-means, but also spectral clustering and agglomerative hierarchical clustering, as they often capture nonlinear distributions better. For nonagglomerative clustering on our model, the number of clusters used was discovered using the elbow method on the graph of the silhouette score and the graph of inertia of the clusters. Since the distribution of points in the latent spaces are somewhat random, the number of clusters, $k$, needed in the analysis are also randomly distributed. However, as displayed in figure 3.8, the elbow point does not move

significantly across the various spaces of varying dimensionality produced by our autoencoder, suggesting that $k = 4$ is in most cases a good choice for the $K$-means algorithm.

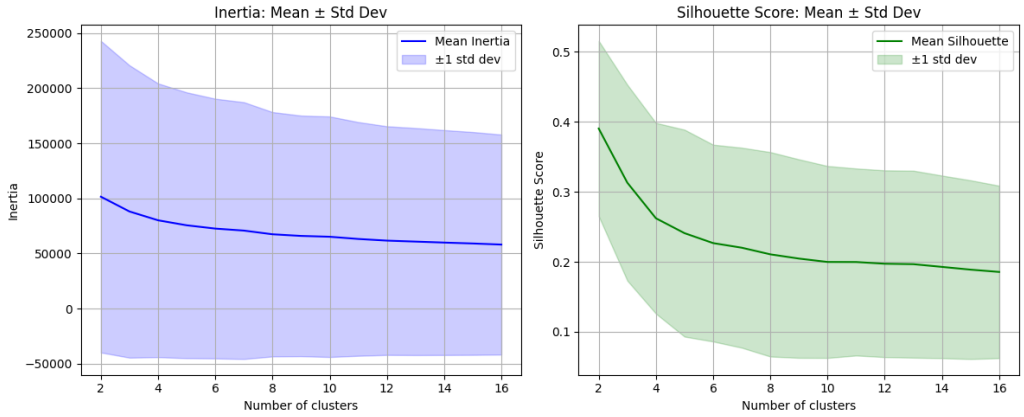



**Figure 3.8:** Deciding the parameter $k$ for $K$-means can be tricky, however we see that the overall shape of the curves are fairly consistent, making the elbow method of finding a good silhouette score fairly consistent. The inflection point lies around $k = 4$ clusters both by using the inertia graph and the graph of the silhouette score, and represents a potential good for the number of clusters.

After clustering the latent space points, a significance analysis was performed, using ANOVA for significant features between clusters and using permutation feature importance to analyze the strength of relationships between clusters.

## 3.4 Existing Baselines

Four baselines were considered when comparing the results of models:

- VAE from Duy Phuong's repository on GitHub [34]
- AE outlined in an article by Shui-Hua Wang et al. [41]
- AE outlined in an article by Hiroyuki Yamaguchi et al. [20]
- AE outlined in an article by Finn Behrendt et al. [13]

To enable efficient implementation of these models, we modified our architecture to allow any model to be built using the same class. This was achieved by

making the encoder and decoder layers, model type (AE or VAE), loss function, and other parameters be configurable inputs. For each model, we fixed the latent space to 128 dimensions and set the $\beta$ value for the VAEs to 1. The following metrics were used to assess model performance.

- Reconstruction loss such as SSIM on both the CN, MCI and AD set.
- Clustering metrics such as Silhouette, Inertia, ARI and clustering accuracy.

Manual inspection of model performance was done utilizing t-SNE, UMAP, spectral clustering and hierarchical clustering. Details about the metrics and inspection methods chosen are explained below.

Image reconstructions were assessed using SSIM due to it being less sensitive to small pixel errors than MSE and focusing more on luminance, contrast and structure which is closer to human perceptual relevance. This makes SSIM a more clinically relevant metric, particularly in scenarios where medical professionals may rely on reconstruction quality to support diagnostic decisions. As for clustering metrics, Silhouette and inertia focuses on grouping performance, whereas ARI and labeling accuracy is additional information that can be useful. Even though this is not a classification task and accuracy metrics are not what we optimize for, labeling accuracy has its purposes. A higher Accuracy or ARI indicates that the latent space has been structured meaningfully, successfully separating distinct clinical groups without direct supervision. This can mean that the learned latent features capture relevant pathological patterns, making the model potentially effective for downstream tasks like anomaly detection. One example of this is if clustering is almost perfect. This could highlight potentially misclassified samples. These could appear as points that deviate significantly from the cluster associated with their label. This sample would then be considered an anomaly relative to that cluster. Besides that, it can help maintain an overview which can help draw connections and conclusions in the future.

Spectral clustering with three clusters was applied to the latent space with nearest neighbors affinity. The number three here is chosen as a standard value due to the dataset being divided into three classes. If one was to assume there are subgroups within any of the datasets, this may not give an accurate picture, however the initial results still provided good insight into the questions about the effect of training on the two reconstruction loss functions and which autoencoder

models are worth pursuing. Upon manual inspection of one of the models, the nearest neighbors parameter (n-neighbors) was set to 25 instead of the default 10 as any number below 20 clustered everything except a few points into one single cluster. This is a parameter that may change or be different for each model, but again, since everything in each model was kept the same except the training loss function and if it should use the AE or VAE method for the latent space, it still provided the insight needed for comparison.

UMAP and t-SNE were used to reduce the latent space to two dimensions for visualization purposes. Since UMAP and t-SNE have their own strengths and weaknesses, both were utilized as a method of cross-validating patterns and clusters which can give a more complete picture of the data.

For spectral clustering, nearest neighbors affinity was used instead of radial basis function (RBF) because it provided better results despite modifying the gamma parameter. Figure 3.9 shows a visual example of why spectral clustering utilizing RBF affinity does not work on the model inspired by Hiroyuki Yamaguchi et al. [20]

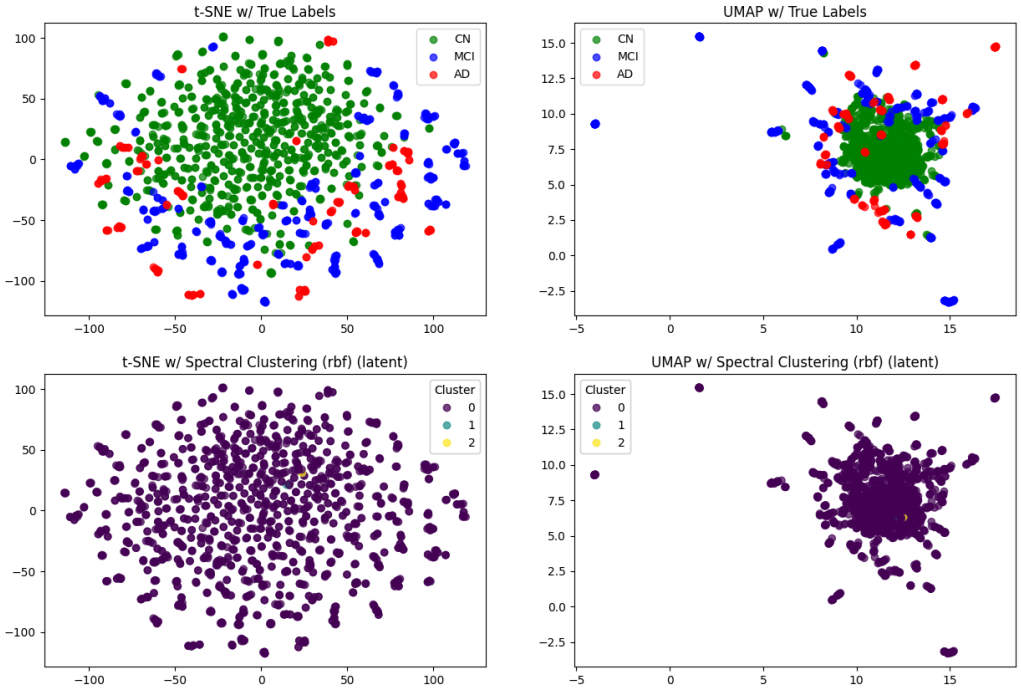



**Figure 3.9:** Spectral clustering with radial basis function (RBF) affinity on the latent space for any gamma value produced similarly unsatisfactory results, which is why we ended up using neareast neighbors affinity. The first two plots show the t-SNE and UMAP projections of the latent spaces corresponding to correctly labeled brain data. This reference shows that the the spectral clustering algorithm is not able to cluster the points into the three clusters, but rather groups the wast majority into group 0 shown as purple circles. Cluster 0, 1 and 2 are arbitrary labels assigned by the spectral clustering algorithm, as it has no access to the true class labels or the underlying meaning of each cluster. **Participant groups:** Green circle — CN. Blue triangle — MCI. Red pentagon — AD.

The final model results were compiled into a table for comparison. Due to the curse of dimensionality, the silhoette and inertia clustering metrics were computed on the 2D UMAP reduced latent space, while ARI and cluster accuracy metrics were calculated based on labels obtained from spectral clustering applied to the latent space. Because both spectral clustering and UMAP are stochastic, the results were computed 10 times per model. The resulting clustering metrics were averaged and presented in a table together with the standard deviation to enable comparison across models.

Besides UMAP and t-SNE for manual inspection, hierarchical clustering was also applied to the latent space such that we could gain a deeper insight into the

clustering without defining a predefined number of clusters.

Given the scope and time constraints of a bachelor's degree. To enable efficient model development and evaluation, a proxy strategy was employed by extracting the middle slice of each brain volume, resulting in 2D representations. This approach significantly reduced computational complexity and allowed for faster experimentation, model iteration, and debugging during the initial phases. Promising models identified through this process were subsequently handpicked and modified for 3D input to facilitate more direct comparisons.

# Chapter 4

# Results

In this chapter, the results of the methods described in Chapter 3 are presented. Section 4.1 shows the behavior of the public ResNet VAE model borrowed from Phuong's repository [34]. Section 4.2 presents the reconstruction loss in different diagnostic categories. In Section 4.3, we examine how different latent space dimensionalities affect model performance. Section 4.4 contains the results of the clustering analysis, including subgroup patterns. Finally, Section 4.5 provides comparative metrics for standard autoencoders (AEs) versus variational autoencoders (VAEs). All results can be found on our GitHub.

## 4.1 Residual Networks Model

The architecture outlined in Phuong's repository [34] seemed promising at first, due to the minimal modifications required for a working implementation on the dataset. One key modification needed was to downscale the input images to half the original size, to reduce training time, as the model is quite big.

The reconstructions after training were disappointing. Figure 4.1 shows how the model fails to capture the variance of the input images and reconstructs a smooth brain-shaped figure. Regardless of learning rate, weight decay and latent space dimensionality the reconstructions were smooth, indicating a big loss in variance.

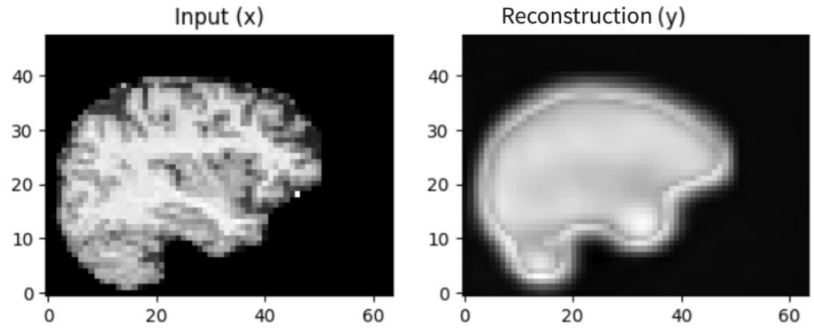



**Figure 4.1:** Prediction by ResNet model of a single individual. The reconstructed image becomes a blurry mean of all the brains in the dataset.

The lost variance in the reconstructions was also reflected in the clustering performance of the latent space produced by the model. Table 4.2 shows how the use of VAEs to encode the images severely reduce clustering performance, as well as reconstruction performance. This suggests that the model is suffering from posterior collapse [47], a phenomenon in which the latent variables becomes uninformative and the decoder largely ignores them under reconstruction. This usually occurs when the posterior $q_\phi(z|x)$ carries limited usable signal, often due to excessive noise or an overly dominant KL term in the loss function. As a result, the decoder produces generic reconstructions, averaging out all training samples. [43]

Although the $\beta$-VAE approach was employed in an attempt to mitigate the collapse, the results did not improve substantially. These limitations, compounded by the fragile structure of the hard-coded repository, ultimately motivated the transition toward building custom model architectures.

## 4.2 Image Reconstructions

Figure 4.2 shows saliency maps, wherein the color of a pixel displays the mean MSE across all images for the corresponding voxel in the median sagittal and transverse planes in the brain. The figure shows how reconstruction loss is higher in the MCI and AD groups, and that the excess reconstruction loss falls in predictable places, aggregating on the surface of the brain and ventricles, indicating brain atrophy. This is in line with earlier findings that brain shrinkage (atrophy) occurs throughout the development of Alzheimer's disease and cognitive impair-

ment. [48]

Furthermore in the figure, a small outline around the brain can be seen, possibly the skull which was not fully removed in preprocessing, suggesting the skull stripping algorithm was too *kind* to the brains, leaving more of the skull than it should. This will likely not cause any changes to the analysis and can safely be ignored.

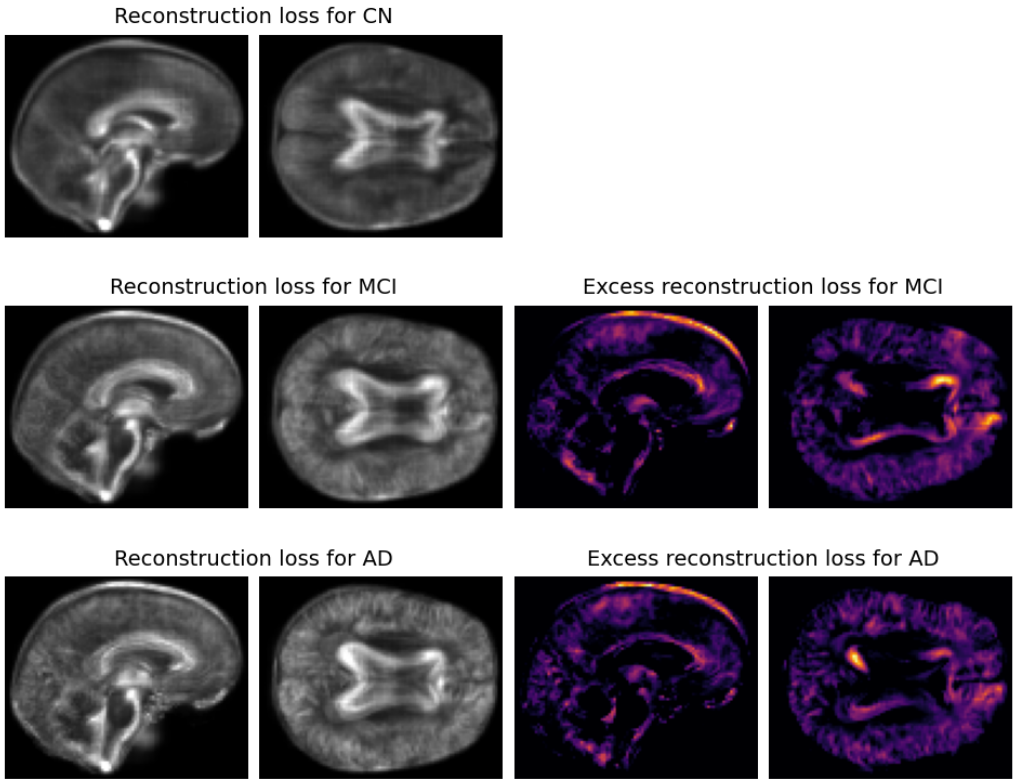



**Figure 4.2:** Mean of reconstruction loss over all images. Models tend to have high reconstruction loss around ventricles and the corpus callosum, regardless of which group the image is sampled from. However, reconstruction loss is higher (ref. excess reconstruction loss) for subjects not from the CN group, particularly in the outermost regions and around the ventricles, indicating brain atrophy.

While reconstruction loss is higher in the MCI and AD groups, significant overlap exists between the performance in each group, as can be seen in figure 4.3, indicating that the models may not be sufficiently sensitive to the distinctions between these groups, limiting their usefulness in real world applications.

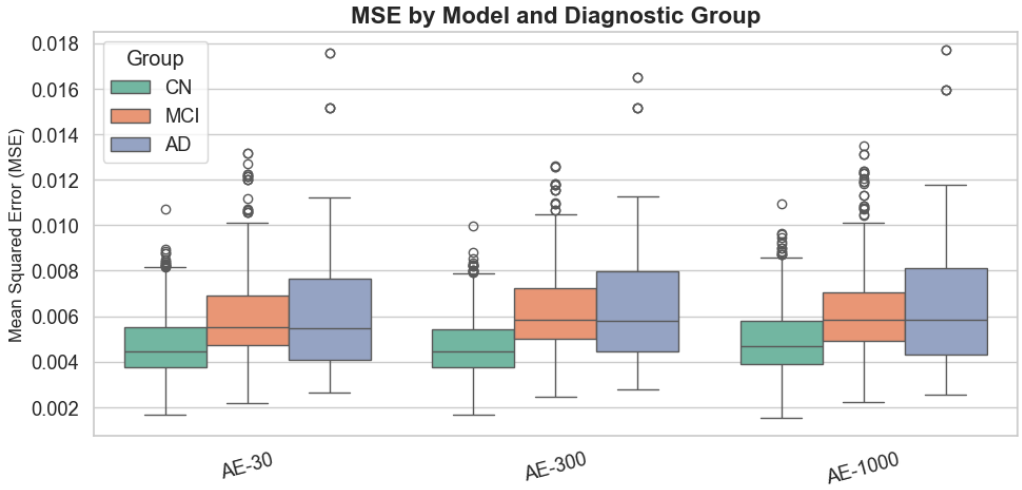



**Figure 4.3:** Performance of three AE models with varying latent space dimensionality. The number of latent dimensions does not seem to impact performance, and while the CN group provides the best results in terms of reconstruction loss, the models provide significant overlap in performance between the groups.

Interestingly, the 95th percentile of the MSE for each image had strikingly similar performance, whereas table 3.2 suggests the 95th percentile should have larger differences in error compared to using the mean loss as reconstruction loss. Figure 4.4 shows how no such improvement in differences was found, providing evidence the models fail to detect anomalies consistently when looking at the diagnostic groups.

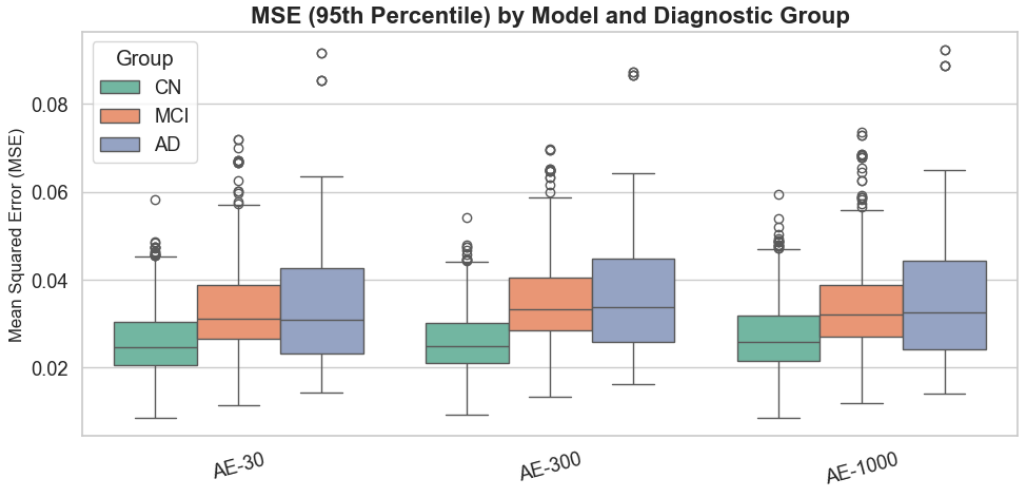



**Figure 4.4:** Performance of three AE models with varying latent space dimensionality. The y-axis shows reconstruction loss using quantile loss (q=0.95) instead of the mean. The overlap between groups is strikingly similar to figure 4.4, using the simple mean loss, with only the relative scale on the y-axis being one order of magnitude larger.

As an anomaly detection algorithm, the autoencoder structure does not perform well enough on the ADNI dataset to confidently differentiate between normal samples and MCI and AD samples. The box plot also shows that the different models have very similar losses for multiple images, as seen in the two outliers in the AD class for all three models, and further shown in the correlation map in figure 4.5. This implies that the models converge to the same loss minimum, regardless of latent space dimensionality.

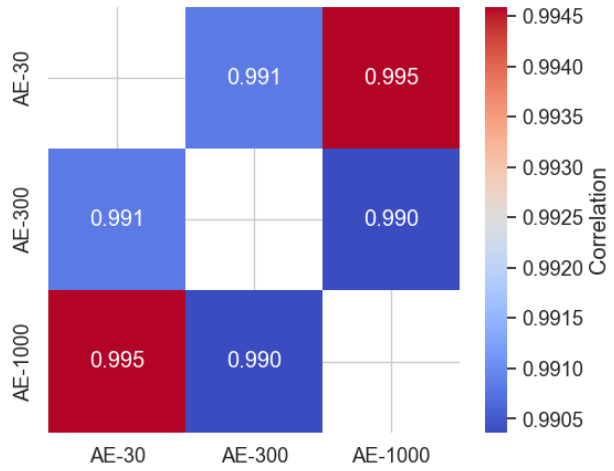



**Figure 4.5:** Loss is highly correlated between models, indicating a common loss minimum is achieved.

## 4.3 Latent Space Encoding

### 4.3.1 Dimensionality

It is of interest to see how many latent dimensions "worth" of data is stored in the input images, to gain insight into how well the models capture meaningful relationships. To find an upper bound estimate for how many dimensions we need to encode the data with little loss, we can employ weak learners to rapidly find an answer. One way of doing this is Principal Component Analysis (PCA), an operation which finds vectors such that a linear combination of n-dimensional inputs per m-dimensional data point minimizes the loss in variance, where n<m. As a linear kernel, PCA is a weak learner and should thus have worse encodings compared to the convolutional autoencoder structure used in this project. Using a convolutional approach to PCA where the model learns to compress sliding windows, thus effectively becoming a convolutional kernel, the weak PCA approach was able to retain at least 98% of variance (the $R^2$) in each of the models four downscaling layers with these numbers of channels or principal components: [4, 12, 36, 108]. By taking the spatially compressed image of shape (4, 5, 4, 108) as the latent space, a total of 8640 latent dimensions are implied. Figure 4.6 compares the PCA reconstruction to the input image, and shows that almost all the

information is retained, however the upscaling process produces ugly “linear” artifacts, resulting in “seams” in the reconstruction, but that is not important for this analysis.

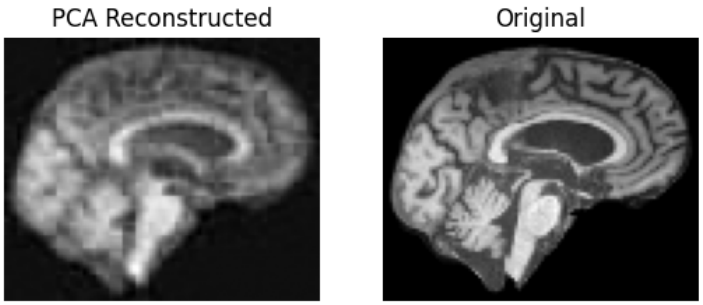



**Figure 4.6:** Slice of an input 3d image and its reconstruction from being encoded using PCA. PCA was applied in a convolutional manner to the image in 4 steps.

Thus we have solid grounds to say that around 10,000 dimensions are more than sufficient to encode the images. Furthermore, by using [4, 12, 36, 108] as the downscaling channels for the autoencoder, we are guaranteed a 98% retained variance.

By changing the number of latent dimensions in the bottleneck layer, we can observe how the reconstruction loss changes. This provides insight into how many latent dimensions are necessary to capture a good encoding of the images. Figure 4.7 shows the loss for the same model trained with different latent dimensions. In particular this model had downscaling layers with these numbers of channels: [8, 16, 32, 64]. After four resolution halvings the image of shape (80, 96, 80) is reduced to (5, 6, 5), thus representing a spatial compression of 1:4096, and a channel depth increase from 1 to 64 channels, giving an information compression rate of 99.61%. A latent space of $5 \cdot 6 \cdot 5 = 150$ dimensions represents the information breakeven point with the downscaling output, meaning above 149 latent dimensions, no further compression occurs for this model.

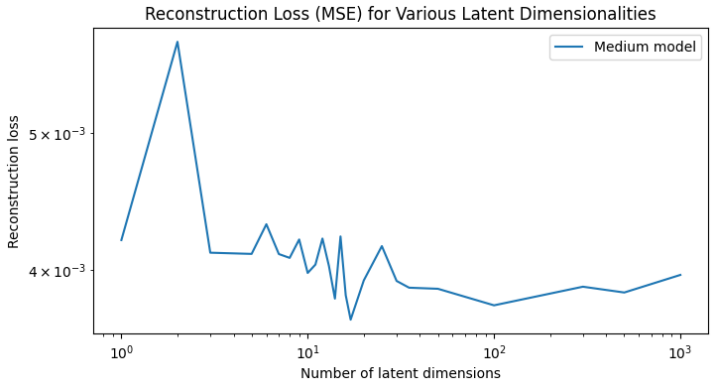



**Figure 4.7:** Reconstruction loss as a function of the number of latent dimensions for this specific architecture, denoted as a *medium model*.

We see that below 10 latent dimensions, the compression is much more unstable, however in some cases the optimizer still manages to provide a low reconstruction loss. After 10 dimensions, no significant gain is achieved in terms of reconstruction. This is very few dimensions, which could either signify very densely packed information, or that little information needs to be encoded since much has already been lost in the downscaling process when extracting spatial information.

### 4.3.2 Information Loss in Downscaling

As seen by applying PCA, depending on the model architecture, simply downscaling the image can result in heavy loss in information. One obvious caveat of exploring the latent space dimensionality this way is that each model architecture will require different numbers of latent dimensions to encode the information that has been downscaled, thus the plot for reconstruction loss versus number of dimensions will look different for each model. To see if too much information is lost in the downscaling process, we can compare the graph of reconstruction loss for models with more parameters in the downscaling layers (large models) with the loss graphs of small models. The model sizes used in this study is summarized in table 4.1, in terms of paramter count.

**Table 4.1:** Parameter count for small, medium and large models in this context. The parameter count includes the encoding and decoding subunits.

| Quantile | Medium | Large | Small | Dimensions |
|---|---|---|---|---|
| 0.25 | 419k | 40M | 56k | 8 |
| 0.50 | 534k | 42M | 85k | 14 |
| 0.75 | 842k | 47M | 162k | 30 |

As visualized in figure 4.8, we found that larger models in general tend to perform worse than their small model counterparts, likely caused by a lack of training time: bigger models need more epochs to minimize loss, however, all models trained were given the same amount of epochs to train on (30 epochs). In the same graph, we also see that in general, models with fewer latent dimensions tend to have a more smooth loss graph, indicating better convergence. The figure shows both large and medium models have much noise in the low-dimensional latent spaces, indicating more training is necessary to achieve convergence.

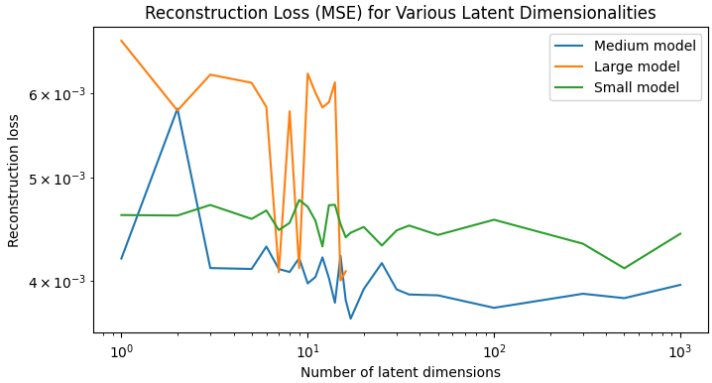



**Figure 4.8:** Large models tend to perform worse after 30 epochs of training than smaller models. Small models tend to perform worse no matter the dimensionality, except for very low-dimensional latent spaces.

Given more time, it would make sense to train the large models for more epochs to try to achieve more stable minimums. No conclusion can be achieved regarding the information loss in the downscaling process, but figure 4.8 showing the graph of loss as a function of number of latent dimensions shows more epochs are needed to achieve convergence.

## 4.4 Clustering

By manual inspection of several of the latent space projections, we see some evidence of clustering in higher dimensions. Applying K-means with k=4 we achieve a silhouette score in n-d of 0.27 (95% CI: [0.20, 0.33]) in n-d, indicating some clustering, but with a somewhat weak structure. [22] However, as silhouette score is based on the euclidean distance between clusters, increasing dimensionality of the space used for clustering lowers the silhouette score, making it hard to gauge performance on the clustering of models of different latent dimensionality.

While the elbow method elected $k = 4$ as the inflection point of the silhouette graph and therefore the best number of clusters to use for $k$-means, experimentation with other values of $k$ showed that having a more relaxed $k$ variable between 4 and 14 clusters gave more interesting results. As shown in figure 4.9, using more clusters gave rise to more significant (using ANOVA) Alzheimer's-related trackers, such as the APOE-4 genotype, indicating more clusters are needed for the level of precision needed in this clustering task. In this figure, an unrelated physical examination of the chest is used to compare the effect of increasing number of clusters, and we see that the Alzheimer's features grow more in significance.

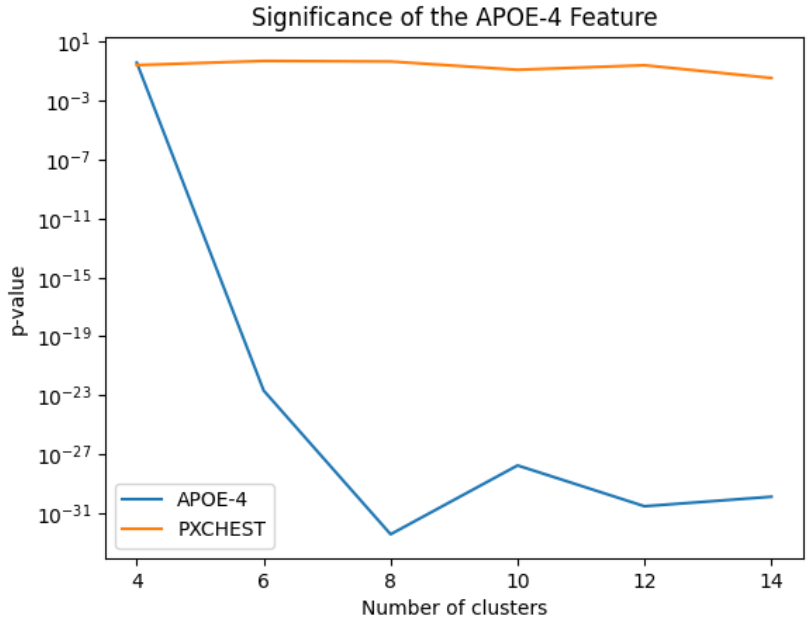



**Figure 4.9:** $p$-value by ANOVA for the APOE-4 genotype for various numbers of clusters, $k$. APOE-4 becomes more significant as the number of clusters increase. Other features, such as PXCHEST (physical examination of the chest) do not grow as rapidly in significance as the Alzheimer's related APOE-4 feature does.

Using ANOVA on the various clustering hyperparameters and taking the average mean $p$-value per feature (as opposed to taking the product of $p$-values, or the likelihood) the most significant features identified were num_events (the number of times a participant was hospitalized for an adverse event), MMSE_MEAN (their average performance on the Mini Mental State Examination), PTDOBYY (year of birth), PXNECK (anomalies in neck region), age, PTEDUCAT (how long a participant has stayed in education), PTGENDER (indicator feature with $1$ being male and $2$ being female), and APOE-4 (counter feature for apoe-4 alleles). A summary of the relative importance of the significant ($p < 0.01$) features is shown in figure 4.10. See appendix B for explanation of ADNI features and their names.

Statistical significance does not indicate the importance in the clustering process of each feature. For this we can use permutation feature importance. Since many of the ADNI features are correlated, we must also group the features so that we can permute each group as a whole. In figure 4.11 grouped permutation feature importance has been performed, using accuracy as the metric of interest on

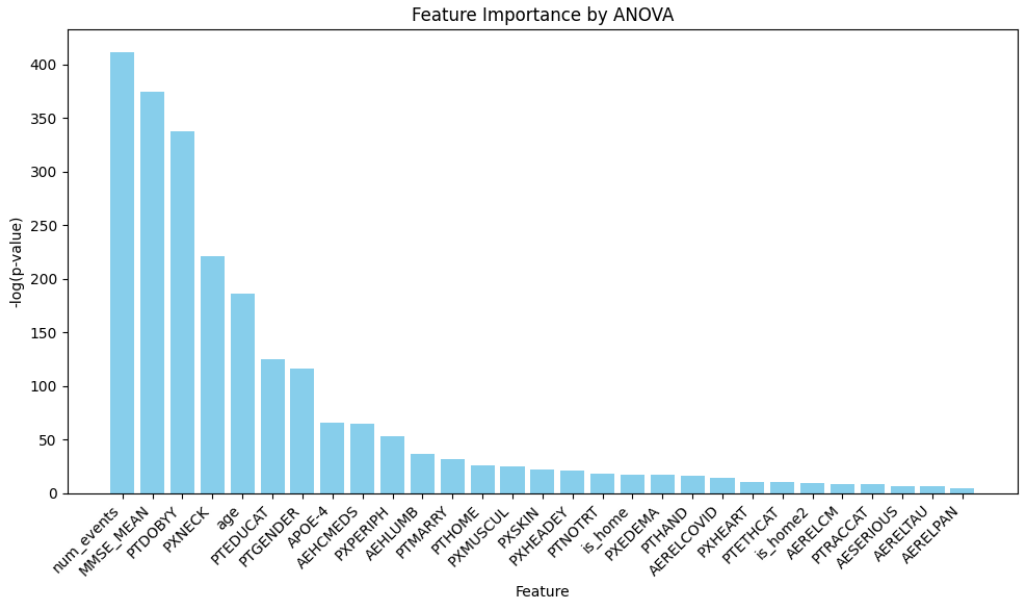



**Figure 4.10:** Feature importance using ANOVA. Most significant features include the number of adverse events a participant has undergone, their MMSE score, and the year they were born.

a random forest trained on the clustering labels over 100 permutation iterations. The permutation groups include Medications, PX features, APOE genotype (both alleles), Age (accouting for date of birth as well as the age feature), Home environment, Brain Volume (calculated as concave and convex), MMSE scores and Adverse events. In the figure we see that medications is now the most important feature to the random forest, in contrast to the overview given by ANOVA in figure 4.10, where each individual medication by itself did not significantly vary between clusters.

It seems strange that some features that seemed highly important using ANOVA are disregarded using permutation feature importance, such as PTGENDER, PTEDUCAT and PX features. This could be due to correlated features *masking* the importance of similar features that convey the same information.

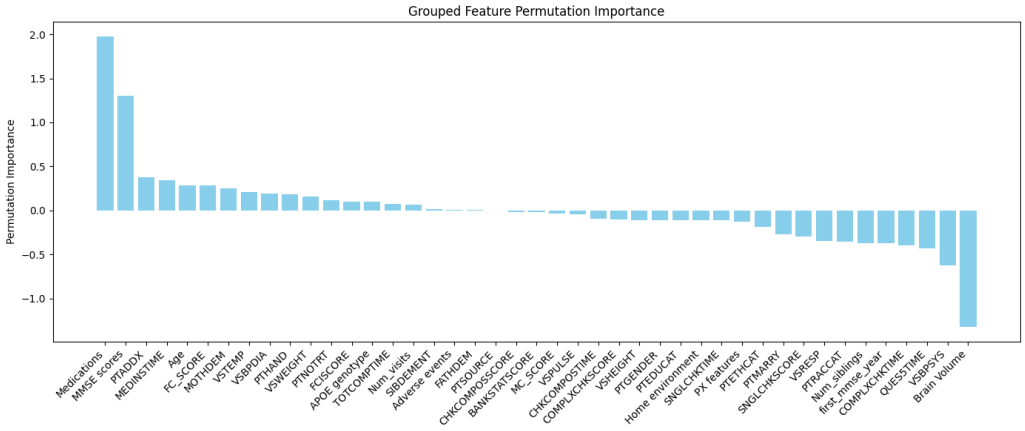



**Figure 4.11:** Permutation feature importance. By grouping features together, medications has become a large contributor to the clustering algorithm.

Figure 4.12 lists features that are significantly ($p < 0.01$) different from cluster to cluster, in order of significance, with the top being the most significant feature. In both trials the most significant feature is cluster_id, referring to the actual cluster each participant belongs to. Increasing the number of clusters we see that the various clusters also exhibit large differences in MMSE (MMSE_MEAN) and number of medically related events (num_events), indicating that more clusters are needed to capture the differences between diagnostic groups when using K-means.

One interesting feature that grew a lot in significance as the number of clusters went from 4 to 15 is PXNECK, a binary signal describing whether the neck is abnormal or normal. The importance of this feature may indicate that the preprocessing steps do not successfully ignore other bodily features of the MRI scans.

Also of interest is the APOE-4 feature, which counts occurrences of the APOE-4 allele in a participant. By increasing the number of clusters to $k = 15$ this feature significantly deviates between clusters.

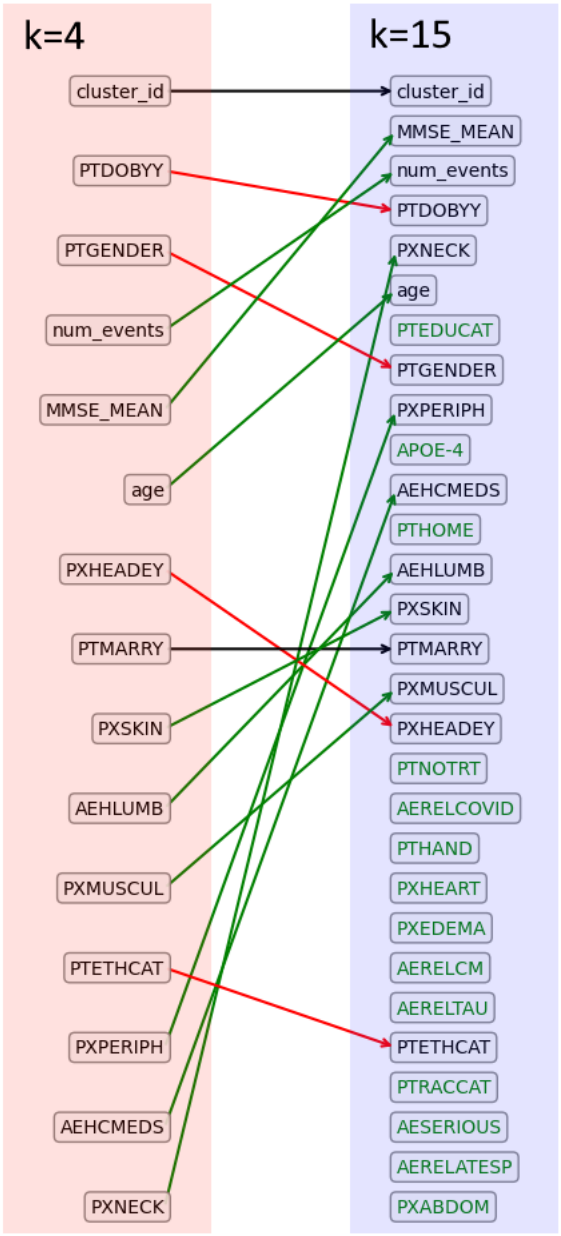



**Figure 4.12:** Features impacting clustering ordered by level of significance (p-value). Increasing the number of clusters from $k = 4$ to $k = 15$ we see many more Alzheimer's related features going up in rank, and many new features appear such as APOE-4 (counter feature for alleles of APOE-4) as well as PXNECK, which may indicate improper pre-processing.

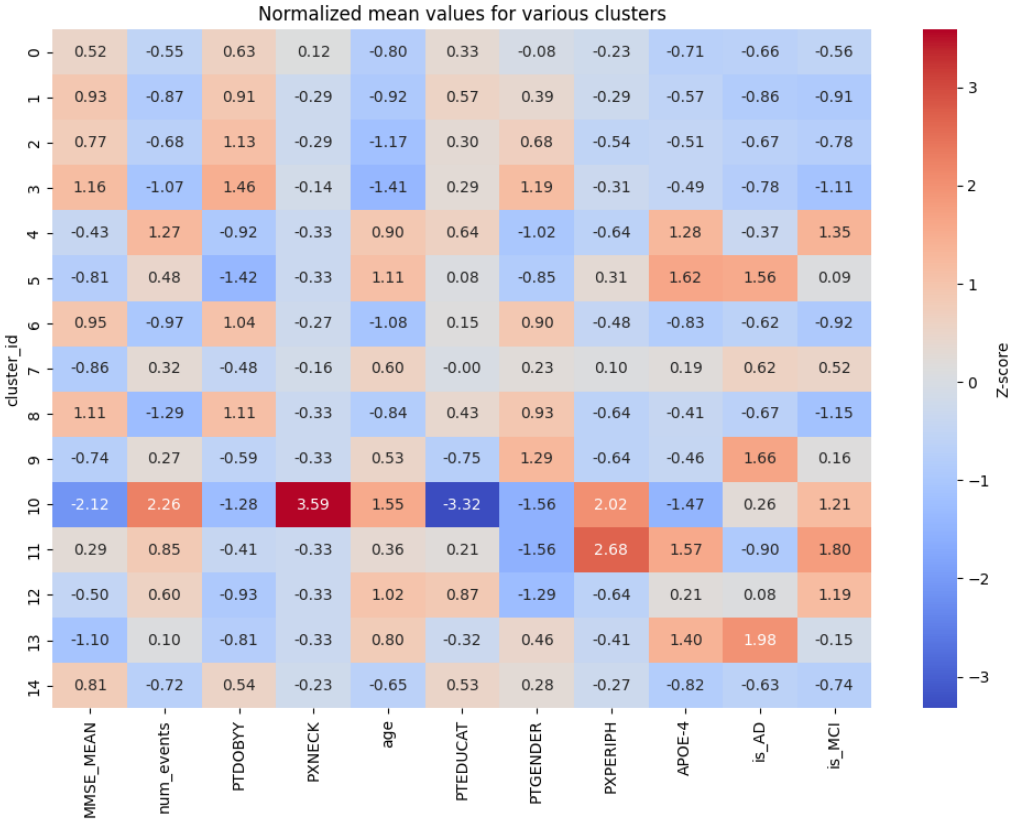



**Figure 4.13:** Heatmap of z-score values for the nine most significant features in the clusters, as well as the AD and MCI rates in each cluster. We notice three distinct AD clusters with ids 5, 9 and 13. The cluster with id 10 significantly deviates in every way measured, and may be considered outliers.

By examining the content of each cluster and how the clusters vary along these features it is possible to find subgroups in the population in which each participant is similar to the rest of the cluster. Figure 4.13 summarizes how the values of these features vary in each cluster, presented as a Z-score.

The figure clearly displays three distinct AD clusters with ids 5, 9 and 13. Examining these clusters by themselves, as displayed in figure 4.14 we see that cluster 9 significantly differs from 5 and 13 with a higher ratio of female participants (PTGENDER), and a lower rate of the APOE-4 allele, highlighting the importance of gender-specific research in the context of AD. Furthermore, cluster 5 has a much earlier date of birth (PTDOBYY) than the other two clusters and is mainly made up of males (PTGENDER). Cluster 13 shares many of the features the other two clusters have and many of the signs associated with a high AD risk such as a high APOE-4 prevalence and a slightly skewed gender distribution.

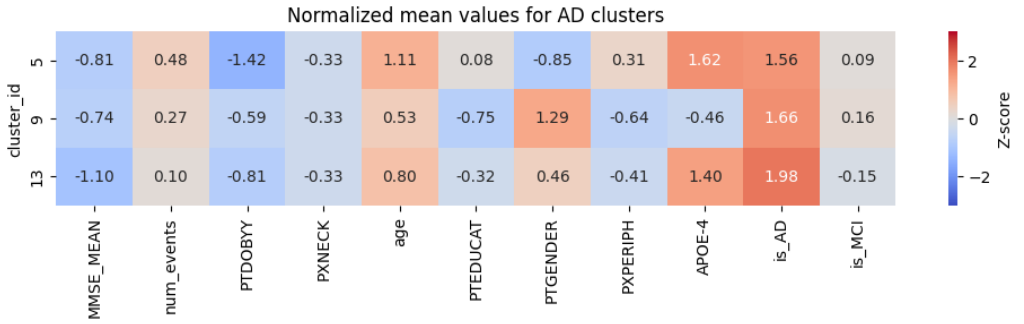



**Figure 4.14:** Selected clusters with significantly high AD prevalence. Cluster 5 has a significantly earlier date of birth (PTDOBYY) and is mainly made up of males (PTGENDER); cluster 9 has a much higher ratio of females and a low APOE-4 prevalence; cluster 13 is somewhere in between the other clusters, displaying a fairly high APOE-4 prevalence, a slightly skewed gender distribution, and generally older participants.

Looking at the biological makeup of the groups it makes sense that group 13 should be somewhere in between cluster 5 and 9, if the latent space is continuous and encodes participants biological data. Looking at the scatter plot of the t-sne projected latent space into 2d in figure 4.15 the expected behavior emerges; cluster 13 (bottom-center) lies between cluster 5 (bottom-left) and 9 (bottom-right), suggesting the latent space captures many of the interesting details of an MRI scan.

Further differences in the clusters, including the ones listed above:

- **Cluster 5** consists mostly of older males, with a low prevalence of home ownership or renting, a high usage of anti-depressants and a somewhat higher weight than the other clusters;
- **Cluster 9** consists mostly of females living in an owned home, with very few heart anomalies, low APOE-4 prevalence and a high prevalence of adverse events related to AV-1451 (tau) tracer
- **Cluster 13** consists of slightly more females than males, living in an owned home, and has a high heart anomaly prevalence and generally low weight.

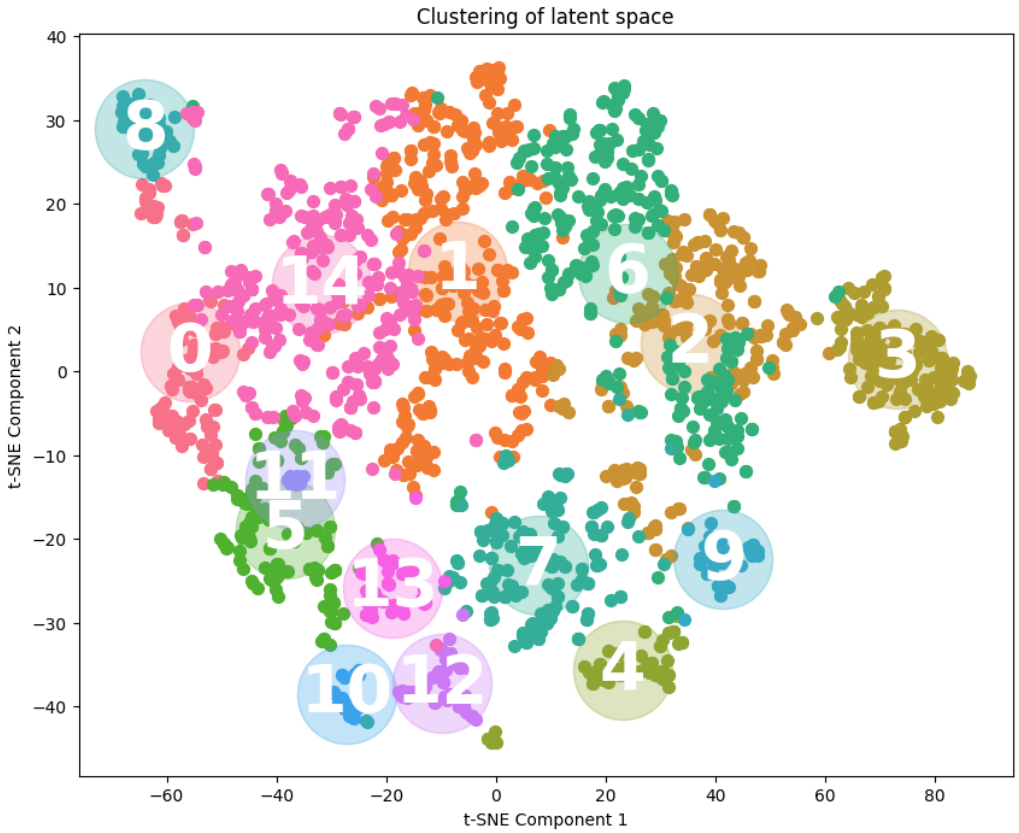



**Figure 4.15:** Clustering of the latent space using $k = 15$ clusters. This can be seen as 15 distinct colored groups in the plot. Notice that some points overlap other clusters due to the clustering happening in the latent space before being projected down to two dimensions (2 components) utilizing t-SNE.

## 4.5 Risk Factors

The features identified to be statistically different between clusters were used on a regression analysis on the probability of an individual belonging to the AD group. The features with the highest impact on probability of AD were gender (PTGENDER), type of home (PTHOME, in particular if an individual belonged to a nursing home or lived in their own house), marital status (PTMARRY), what ratio of their siblings have developed dementia (SIBDEMENT), and status of peripheral vascular (PXPERIPH). These findings support recent articles finding a higher risk of AD among married individuals compared to other marital status groups [40] and among females compared to men [9].

## 4.6 Comparison to Existing Baselines

The first experiment for the modified architecture was inspired by three research papers and a public GitHub repository, resulting in 4 distinct architectures. [34, 41, 20, 13] The primary objective was to compare the performance of VAEs and standard AEs in terms of reconstruction loss and clustering quality. Additionally, two loss functions ($\mathcal{L}$: SSIM vs MSE) for training were tested to assess their impact on model performance. This resulted in 16 models in total. Table 4.2 shows the result of the 2D models tested.

| **Model Information** | | | **SSIM** | | | **Clustering Metrics** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Author | $\mathcal{L}$ | Type | CN | MCI | AD | Silhouette | Inertia | ARI | Cluster accuracy |
| Duy Phuong [34] | MSE | AE | 0.79 | 0.64 | 0.62 | 0.24 ± 0.01 | 15 552 ± 1 432 | 0.03 ± 0.00 | 0.45 ± 0.00 |
| | SSIM | AE | 0.79 | 0.62 | 0.62 | 0.29 ± 0.01 | 17 895 ± 826 | 0.20 ± 0.00 | 0.62 ± 0.00 |
| | MSE | VAE | 0.56 | 0.44 | 0.46 | -0.01 ± 0.01 | 3 802 ± 125 | -0.00 ± 0.00 | 0.35 ± 0.00 |
| | SSIM | VAE | 0.59 | 0.48 | 0.49 | -0.01 ± 0.01 | **3 729** ± 87 | 0.00 ± 0.00 | 0.36 ± 0.00 |
| Shui-Hua Wang et al. [41] | MSE | AE | 0.73 | 0.59 | 0.57 | 0.32 ± 0.01 | 10 694 ± 172 | 0.06 ± 0.00 | 0.58 ± 0.00 |
| | SSIM | AE | 0.74 | 0.60 | 0.58 | 0.15 ± 0.01 | 17 050 ± 1 149 | 0.00 ± 0.00 | 0.46 ± 0.00 |
| | MSE | VAE | 0.41 | 0.33 | 0.34 | -0.02 ± 0.01 | 3 888 ± 116 | -0.00 ± 0.00 | 0.34 ± 0.00 |
| | SSIM | VAE | 0.35 | 0.31 | 0.33 | -0.02 ± 0.00 | 3 795 ± 52 | 0.00 ± 0.00 | 0.35 ± 0.00 |
| Hiroyuki Yamaguchi et al. [20] | MSE | AE | **0.85** | **0.68** | **0.66** | 0.28 ± 0.01 | 15 456 ± 1 947 | 0.27 ± 0.00 | 0.64 ± 0.00 |
| | SSIM | AE | 0.84 | 0.65 | 0.65 | 0.35 ± 0.01 | 18 532 ± 1 513 | 0.30 ± 0.00 | 0.72 ± 0.00 |
| | MSE | VAE | 0.55 | 0.45 | 0.45 | -0.01 ± 0.01 | 3 683 ± 92 | -0.00 ± 0.00 | 0.35 ± 0.00 |
| | SSIM | VAE | 0.58 | 0.47 | 0.48 | -0.02 ± 0.00 | 3 847 ± 70 | -0.00 ± 0.00 | 0.34 ± 0.00 |
| Finn Behrendt et al. [13] | MSE | AE | 0.76 | 0.65 | 0.64 | 0.34 ± 0.01 | 31 726 ± 2 495 | 0.35 ± 0.00 | 0.70 ± 0.00 |
| | SSIM | AE | 0.75 | 0.62 | 0.62 | **0.38** ± 0.02 | 21 979 ± 1 455 | **0.35** ± 0.00 | **0.73** ± 0.00 |
| | MSE | VAE | 0.49 | 0.39 | 0.40 | -0.02 ± 0.00 | 3 842 ± 65 | -0.00 ± 0.00 | 0.36 ± 0.00 |
| | SSIM | VAE | 0.54 | 0.44 | 0.45 | -0.02 ± 0.01 | 3 772 ± 49 | -0.00 ± 0.00 | 0.35 ± 0.00 |

**Table 4.2:** Evaluation of clustering and reconstruction performance for 4 2D autoencoder models, each tested across 4 settings combining training loss ($\mathcal{L}$: SSIM vs MSE) and model type (AE vs VAE). The best results for each metric are highlighted in bold.

From the table a few conclusions can be drawn. However, some of the results should be interpreted with caution. The reconstruction loss for the healthy brains (CN) does not reflect out-of-sample performance, as K-fold cross-validation has not been applied. This is because the primary objective is not to compare reconstruction loss across models within the CN set. Instead, the focus is on identifying anomalies and deviations in the reconstruction and clustering of MCI and AD samples. That being said, the following conclusions can be drawn.

VAEs perform significantly worse for all metrics across all models, except inertia. This can be seen visually in figure 4.16, showing that VAEs achieves a silhouette centered around 0 indicating no clustering, and inertia significantly lower than their AE counterparts.

The same figure also shows how using SSIM or MSE as the model loss to be backpropagated during training does not significantly impact performance in either clustering or reconstruction metric. Model loss function does also not impact reconstruction performance, as summarized in table 4.3, showing mean SSIM remaining the same down to 2 decimals for either loss function.

**Table 4.3:** Mean SSIM values by loss function for each diagnostic group.

| $\mathcal{L}$ | $\text{SSIM}_{CN}$ | $\text{SSIM}_{MCI}$ | $\text{SSIM}_{AD}$ |
|---|---|---|---|
| MSE | 0.643 | 0.521 | 0.517 |
| SSIM | 0.647 | 0.524 | 0.527 |

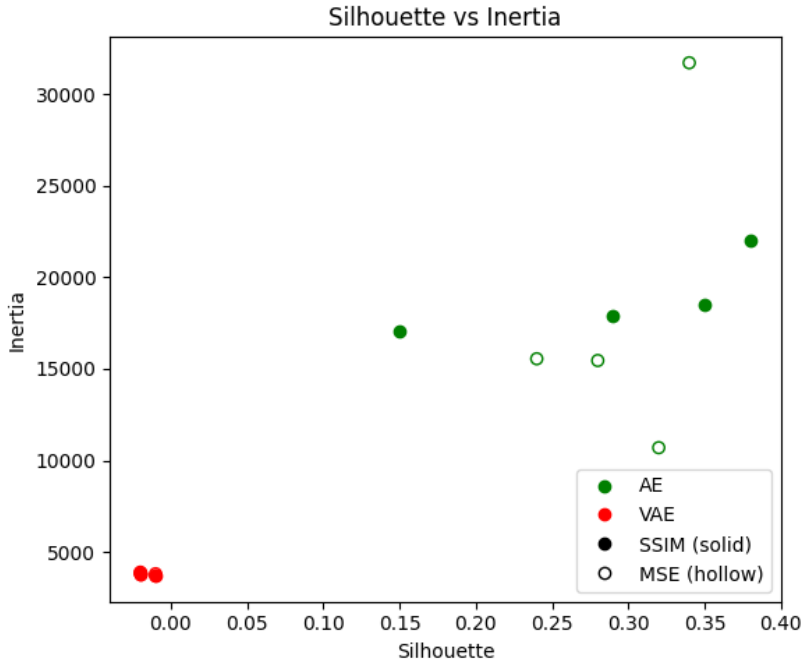



**Figure 4.16:** Scatter plot of Silhouette (x-axis) vs Inertia (y-axis) for the 16 models with some points overlapping. Silhouette scores and inertia are lower for VAEs compared to AEs. No clear relationship with the loss function (SSIM or MSE) and Silhoutte and Inertia is seen. VAEs perform worse measured on silhouette, better when measured on inertia. This indicates worse clustering overall.

The drop in clustering performance by using VAEs as compared to simple AEs is due to the latent space being regulated to be more normally distributed by the KL loss term. In practice, the KL loss term trades precision for generalization by sacrificing structure for smoothness. Although VAEs are good for generative modeling and exploring the data manifold, their design make them less ideal for tasks requiring sharp reconstruction and well-separated latent clusters, such as anomaly detection or disease subtype classification.

Inertia, which measures how tightly grouped the samples in each cluster are to the center, is very low for the VAE models. Low inertia in this context indicates that the model has over smoothed the latent space, causing the representations of different classes (CN, MCI, AD) to overlap heavily. In these tests, the beta parameter (the KL loss regularization strength) of the VAE models was set to 1, lower beta values may perform better than a traditional AE (which by extension has a beta parameter of 0, indicating no KL regularization). Due to time constraints and clustering objectives stated in 1.2, the effect of the beta parameter was not tested beyond 0 and 1. Our assumption is that punishing the latent space for not being normally distributed, blurring boundaries and suppressing outliers is counterproductive in our purposes of anomaly detection.

One other noticeable detail is that the reconstruction error is slightly higher for AD than MCI overall for all non VAE models. This indicates that AD brains indeed do have higher deviations from the normal than MCI does, as was also shown in the earlier figure 4.2

The model by Finn Behrendt et al. [13] had the highest Silhouette score (0.38 $\pm$ 0.02), ARI (0.35 $\pm$ 0.00) and clustering accuracy (0.73 $\pm$ 0.00). To gain a better intuitive understanding of the result, a visual representation of the model results was created. Figure 4.17 shows a visual inspection of the clustering results produced by Finn behrendt el al. [13].

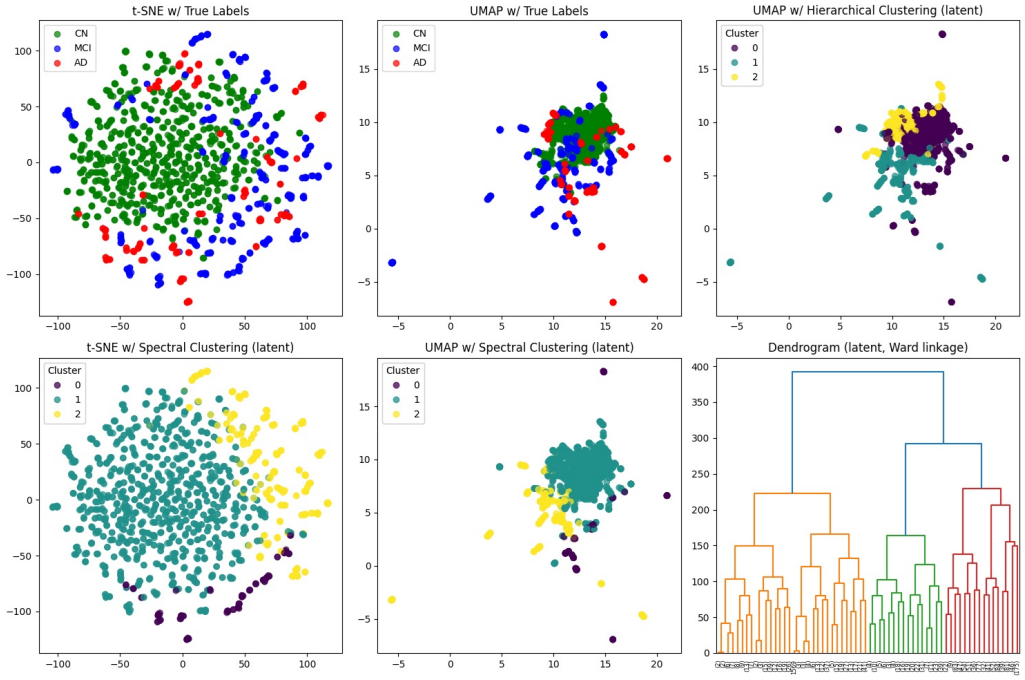



**Figure 4.17:** 2D projection and clustering of the Finn Behrendt et al. [13] architecture. The first two plots (t-SNE w/True Labels and UMAP w/True Labels) show the True colored clusters CN, MCI, AD as colored by green, blue, red. This is used as reference to see if the different diagnoses can be distinguished from each other. The two plots just below are attempts at utilizing spectral clustering on the latent space to see if we can distinguish between different clusters. The two outermost right plots utilize Hierarchical clustering on the latent space. This serves the purpose of visualizing clusters beyond defining a specific number of clusters.

As observed in both the t-SNE and UMAP plot, the majority of the CN group is successfully clustered together with some distance to the MCI and AD set. However the distinction between MCI and AD proves to be more challenging. Spectral clustering on the latent space seems to have made a decent job at seperating the three clusters (CN, MCI, AD) into the three clusters (0, 1, 2). As the spectral clustering doesnt know which number to associate with what class, it just applies a number to the cluster. Moreover, due to the inaccurate labeling produced by hierarchical clustering on this model, no definitive conclusions or further hypothesis can be made on based on the dendrogram results.

Overall, the fact that clusters overlap so much could be the result of insufficient data in the MRI scans to distinguish between the CN, MCI and AD set. This

makes sense as typical diagnosis usually also include biomarkers, cognitive tests and medical history.

The 4 best performing 2D models, which used the standard autoencoder architecture, were selected for conversion to 3D models for comparison. Although there was no significant difference in performance between MSE and SSIM training losses, MSE was used because it trains faster. Furthermore, our own model with 1-4 halvings were handpicked to compare against the existing benchmark. Due to computational time and memory issues, the dataset was reduced by half.

Table 4.4 shows the result of these 8 3D autoencoder models.

| **Model Information** | | | **SSIM** | | | **Clustering Metrics** | | | |
|---|---|---|---|---|---|---|---|---|---|
| Author | $\mathcal{L}$ | Type | CN | MCI | AD | Silhouette | Inertia | ARI | Cluster accuracy |
| Duy Phuong [34] | MSE | AE | 0.17 | 0.10 | 0.11 | **0.50** $\pm$ 0.01 | 7 001 $\pm$ 162 | 0.03 $\pm$ 0.00 | 0.40 $\pm$ 0.00 |
| Shui-Hua Wang et al. [41] | MSE | AE | 0.89 | **0.77** | **0.76** | 0.26 $\pm$ 0.00 | 7 102 $\pm$ 1 003 | -0.00 $\pm$ 0.00 | 0.40 $\pm$ 0.00 |
| Hiroyuki Yamaguchi et al. [20] | MSE | AE | 0.01 | 0.01 | 0.01 | 0.45 $\pm$ 0.03 | 11 675 $\pm$ 674 | 0.01 $\pm$ 0.00 | 0.41 $\pm$ 0.00 |
| Finn Behrendt et al. [13] | MSE | AE | 0.55 | 0.53 | 0.52 | 0.46 $\pm$ 0.01 | 37 591 $\pm$ 3 890 | 0.01 $\pm$ 0.00 | 0.39 $\pm$ 0.00 |
| Our Autoencoder (1 halving) | MSE | AE | **0.91** | 0.64 | 0.66 | 0.49 $\pm$ 0.00 | 13 746 $\pm$ 548 | 0.00 $\pm$ 0.00 | 0.39 $\pm$ 0.00 |
| Our Autoencoder (2 halvings) | MSE | AE | **0.91** | 0.66 | 0.64 | 0.33 $\pm$ 0.00 | 3 879 $\pm$ 230 | 0.03 $\pm$ 0.00 | 0.45 $\pm$ 0.00 |
| Our Autoencoder (3 halvings) | MSE | AE | 0.66 | 0.49 | 0.47 | 0.33 $\pm$ 0.00 | 3 697 $\pm$ 49 | 0.03 $\pm$ 0.00 | 0.45 $\pm$ 0.00 |
| Our Autoencoder (4 halvings) | MSE | AE | 0.77 | 0.68 | 0.67 | 0.39 $\pm$ 0.01 | **2 990** $\pm$ 28 | **0.05** $\pm$ 0.00 | **0.46** $\pm$ 0.00 |

**Table 4.4:** Results of comparing 8 3D autoencoder models with a latent space of 128 across different reconstruction metrics and clustering metrics. The best results for each metric are marked in bold.

While our autoencoder shows stronger reconstruction performance on the CN set, CN results are not validated out-of-sample through K-fold cross-validation. Among the compared models, the one by Shui-Hua Wang et al. appears most promising for image anomaly detection, achieving the highest SSIM scores on MCI and AD subjects. The model by Duy Phuong and our autoencoder with four halvings may be better suited for clustering tasks and identifying individuals who deviate significantly from expected patterns in their given cluster, due to higher silhouette and ARI score. As for clustering it is difficult to say which model seems the most promising. Although the AE by Duy Phuong seems to have the highest silhouette score reaching 0.50 $\pm$ 0.01, it also has an extremely low reconstruction value due to not having been fully trained. Training on this model was stopped early due to it running for several days without converging and could definitely give better and different results if it was ran for longer. The next models that could have potential for clustering is the model by Finn Behrendt et al. or our own model with only 1 halving due to high silhouette or our model with 4 halvings due to decent silhouette, highest inertia (2990 $\pm$ 28), ARI (0.05) and cluster accuracy

(0.46).

Comparison with the 2D models is difficult due to the dataset being halved per class (CN, MCI and AD) from the original set as outlined in table 3.1 due to computational time. Additionally, while increasing the dimensionality of the latent space could improve performance. Further extensive tests of many other custom models (not included in this thesis) suggest that the MRI data alone are insufficent for clearly seperating the diagnostic categories through clustering from the latent space of standard autoencoders. This aligns with the understanding that Alzheimer's diagnosis typically requires more than imaging alone, often incorporating biomarkers, cognitive testing and clinical history.

# Chapter 5

# Discussion

This chapter discusses the implications of the findings, the limitations of the study, and the extent to which the results can be generalized.

## 5.1 Implications

We found several features that differ significantly between clusters of brains. These features do not only apply to anomaly detection but also help to see different *types* of anomalies by providing context. While in most cases these cluster features are biological factors known to cause large differences in the brains of subjects, such as gender and age, the findings imply that also several medical procedures and past events in life affect how AD is expressed by a subject. Based on these factors, more accurate care can potentially be given to prevent rapid development of AD.

Three distinct clusters of AD were identified, differing in gender, age and prevalence of the APOE-4 allele. These findings do not necessarily imply AD can be divided into different disease subgroups altogether, but imply that there are many ways people develop AD, and therefore also many ways to protect against it. Furthermore, future research must take into consideration that study participants may not all belong to the same group or cluster, potentially giving misleading results.

## 5.2 Limited Generalizability

Some features in the ADNI dataset are highly correlated. However, balancing the dataset by removing participants until the correlation is insignificant also reduces the already small dataset beyond what is reasonable. Due to correlated features of the participants in the study, in particular age, MMSE scores and education, it is hard to know the exact reason the results came out as they did. Perhaps clusters form around groups of individuals based on the large effect of all these correlated features, which would inflate the implications of them in the analysis, while in the true population, no such division exists. Therefore, the generalizability of the study must be questioned. Future work should build on this by inspecting the effect of these relationships.

Generalizability must also be further questioned due to the longitudinal nature of the ADNI dataset. While in the true population, diagnoses are continually changing, albeit minutely, in the ADNI dataset, no such change was found. Participants of the AD group belonged to this group from start to end of the study. This may be due to misinterpretation of the study files of the dataset, lack of metadata, or chance.

## 5.3 Limitations

### 5.3.1 Imbalanced Dataset

As seen in the correlation matrix in figure 3.1, two key correlations in the dataset pose potential challenges: age with MMSE score (Mini-Mental State Examination) and education with MMSE score. These relationships may influence the interpretation of cognitive assessments and should be carefully considered in future works. ADNI participants are mostly from the US and Canada, however these correlations are not necessarily true for other populations. These imbalances cause algorithmic bias in the models, potentially explaining why many of the significant features highlighted in the report are non-pathological.

### 5.3.2 Training Time

Training large models takes long to converge. While it is possible to create a single large model and analyze the capabilities of this model, due to the exploratory

nature of this project, and the decire to look at many different model configurations and see how the latent space changes accordingly, very large models were for the most part omitted from our research. With more time and compute, perhaps more interesting latent space representations could be developed. This also caused the reconstructed images of many models to be blurry, since sharp images require more parameters and epochs of training, which may or may not capture the interesting details in the latent space.

### 5.3.3 Longitudinal Data

While many MR images were processed while training these models, we only had a relatively small number of individuals participating in the study, as summarized in table 3.1. This could lead to various unwanted side effects such as subjects who are both imaged in the test and train sets, and overfitting, due to the artificially lowered variance of the data compared to truly independent samples.

### 5.3.4 Outlier Removal

Results could potentially be improved by further refining the preprocessing pipeline by including manual inspection to remove images of patients who do not represent the distribution very well, as they can spread out the models' learned distribution, thus making the model worse at detecting anomalies. In the case of this study, not enough time or background information was in place for removing images that are clinically far from the mean.

# Chapter 6

# Conclusions

This chapter summarizes the key findings from the experimental evaluation of autoencoder based models applied to brain MRI scans. It reflects on the significance of the results, addresses unresolved challenges, and outlines directions for future research aimed at improving model interpretability, reconstruction quality, and anomaly detection.

## 6.1 Experimental Results

The results outlined in this study demonstrate the potential and the limitations of autoencoder based models in the domain of brain MRI scans. By tuning hyperparameters and model architectures we explored how well different models could encode the images into latent spaces. Applying various clustering algorithms to these latent spaces revealed only limited clustering. Our analysis revealed several key findings:

- **Latent Dimensionality:** Models with fewer than 10 latent dimensions exhibited unstable behavior, with inconsistent reconstruction loss. In some cases, the optimizer succeeded in converging to a low loss, suggesting that a small number of latent dimensions can still capture critical features, but final validation loss for the runs varied a lot. Beyond 10 dimensions, additional capacity resulted in diminishing returns, indicating that essential information is highly concentrated in a compact latent space.
- **Reconstruction Loss as a Diagnostic Signal:** Models were able to

highlight images with abnormally high reconstruction error, providing clues about structural anomalies such as atrophy, but with a lot of uncertainty. Using the reconstruction alone as a diagnostic signal is inadvisable, in particular in disease specification, since the overlap between MCI and AD groups were much greater than between the two and the CN group. These reconstruction patterns aligned in many cases with clinical expectations for neurodegenerative conditions.

- **Latent Space Clustering:** Only weak clustering was observed in the latent spaces of the models, suggesting most variables had continuous transitions, instead of having multiple modes. No evidence was found for statistically different modes of MCI or AD in the dataset, but the clustering did also not disprove it. When it comes to biological and historical factors. Clustering analysis revealed that differences in brain structure were influenced not only by known biological variables (e.g., gender, age) but also by past adverse events and medical history. These findings support the view that Alzheimer's Disease (AD), may express itself in different ways, each potentially requiring different protective or therapeutic strategies.
- **Autoencoder Type and Training Loss:** All VAEs significantly under perform the standard autoencoders (AEs) in the reconstruction metric SSIM and the clustering metrics Silhouette, ARI and labeling accuracy except Inertia. This is due to the VAEs over smoothing the latent space, blurring boundaries and suppressing outliers by forcing the latent space to be more normally distributed. Thus, it is inadvisable to choose VAEs over standard AEs for anomaly detection in both image reconstruction and clustering of the latent space. Using SSIM or MSE as the model loss ($\mathcal{L}$) to be backpropagated during training for either VAEs or AEs does not have an observable impact on performance in either of the clustering or reconstruction metrics. As a consequence, training on MSE is advisable as it computes faster.

Together, these results demonstrate that while unsupervised models still face challenges in precisely segmenting disease subtypes, they have potential within anomaly detection, subject stratification, and diagnosis support. They also offer insights into the complex and multifactorial nature of structural brain differences observed in neurodegenerative diseases.

## 6.2 Future Work

### 6.2.1 Blur in Reconstructed Images

The models could never produce sharp images, indicating that many of the smaller details were lost in the encoding. Further work should be done to examine the importance of these details and whether sharper reconstructed images would lead to better clustering in the latent space. The blur could potentially be minimized by adding more parameters to the models, however, as seen in this project, adding more parameters necessitates a bigger dataset and more epochs to converge. If the blur cannot be removed from the reconstructed images this way, it would also be of interest to look at lower resolution images, as they would require fewer parameters to reproduce with high fidelity, thus allowing for sharper reconstructions.

### 6.2.2 Data Augmentation

This project used convolutional models for encoding images into latent representations, however convolutional models are known to perform better when data is augmented to provide more data than the model otherwise would have been trained on. In this project, no augmentation was performed in the preprocessing step, thus potentially removing these gains in performance.

### 6.2.3 Multimodal Models

As we have shown, models trained only on MR images using an unsupervised approach struggle to differentiate between types of anomalies, in particular MCI and AD subjects. To improve differentiation between anomaly types, more modes of information could be added, such as basic biological information or PET scans. While the intersection between publicly available recorded PET scans and MRI scans is small, using a GAN or a similar architecture could provide a more detailed insight into the makeup of a subject's brain. One idea that struck us during model testing was that perhaps one way to find the best model for any architecture be it AEs, VAEs, GANs, etc. Would be to utilize the same proxy strategy of optimizing 2D models for speed, to weed out the best solutions that can be converted to 3D models. Automation of model builds could then be achieved by utilizing genetic

algorithms for the search space reduction in deciding which layer types and combinations should be fed into the autoencoder class as input parameters. To enable a more human breadth first search, a parameter for the layer depth (number of layers for the encoder and decoder) could be incremented every n generation. Rewarding population diversity in the fitness function could potentially also be useful to encourage exploration.

# Appendix A

# GitHub

All code and results used in this thesis, except ADNI data are publicly available on GitHub: https://github.com/jon-tj/DementiaMRI.

# Appendix B

# Abbreviations of Features in the ADNI Dataset

Below is a brief overview of the features in the ADNI dataset that were found to be significant, along with a description of what they measure and which table they were fetched from.

- **AEHCMEDS:** (Adverse events) Concurrent Medication Prescribed or Changed
- **AEHLUMB:** (Adverse events) Related to Lumbar Puncture
- **AERELATSP:** (Adverse events) Related to other study procedure(s)
- **AERELCM:** (Adverse events) Related to concomitant therapy
- **AERELCOVID:** (Adverse events) Related to COVID-19 illness
- **AERELPAN:** (Adverse events) Related to COVID-19 pandemic disruption
- **AERELTAU:** (Adverse events) Related to AV-1451 (Tau) tracer
- **AESERIOUS:** (Adverse events) Was event serious at any time during the trial?
- **BANKSTATSCORE:** (Financial capacity) Bank Statement Management Score
- **CHKCOMPOSSCORE:** (Financial capacity) Combined Checkbook/Register Score

- **CHKCOMPOSTIME:** (Financial capacity) Checkbook/Register Composite Time
- **COMPLXCHKSCORE:** (Financial capacity) Complex Checkbook/Register Score
- **COMPLXCHKTIME:** (Financial capacity) Complex Checkbook/Register Time
- **FATHDEM:** (Family history) Did/Does the biological father have dementia?
- **FCISCORE:** (Financial capacity) TOTAL FCI-SF SCORE
- **FC_SCORE:** (Financial capacity) Financial Conceptual Knowledge Score
- **MC_SCORE:** (Financial capacity) Mental Calculation Score
- **MEDINSTIME:** (Financial capacity) Time to complete Item 4 (Health Care Insurance Problem)
- **MOTHDEM:** (Family history) Did/Does the biological mother have dementia?
- **PTADDX:** (Demographics) Year of Alzheimer's Disease diagnosis
- **PTDOBYY:** (Demographics) Participant Year of Birth
- **PTEDUCAT:** (Demographics) Participant Education
- **PTETHCAT:** (Demographics) Ethnic Category
- **PTGENDER:** (Demographics) Participant Gender
- **PTHAND:** (Demographics) Participant Handedness
- **PTHOME:** (Demographics) Type of Participant residence
- **PTMARRY:** (Demographics) Participant Marital Status
- **PTNOTRT:** (Demographics) Participant Retired?
- **PTRACCAT:** (Demographics) Racial Categories

- **PTSOURCE:** (Demographics) Information Source
- **PXABDOM:** (Physical examination) Abdomen
- **PXEDEMA:** (Physical examination) Edema
- **PXHEADEY:** (Physical examination) Head, Eyes, Ears, Nose and Throat
- **PXHEART:** (Physical examination) Heart
- **PXMUSCUL:** (Physical examination) Musculoskeletal
- **PXNECK:** (Physical examination) Neck
- **PXPERIPH:** (Physical examination) Peripheral Vascular
- **PXSKIN:** (Physical examination) Skin
- **QUES5TIME:** (Financial capacity) Time to complete Item 5 ( Tax Credit)
- **SIBDEMENT:** (Family history) Did/Does this sibling have dementia? This feature has been aggregated using the mean of siblings.
- **SNGLCHKSCORE:** (Financial capacity) Single/Checkbook Register Score
- **SINGLCHKTIME:** (Financial capacity) Time to complete items 7-16
- **TOTCOMPTIME:** (Financial capacity) TOTAL Composite Time
- **VSBPDIA:** (Vital signs) Diastolic - mmHg
- **VSBPSYS:** (Vital signs) Systolic - mmHg
- **VSHEIGHT:** (Vital signs) Height
- **VSPULSE:** (Vital signs) Seated Pulse Rate (per minute)
- **VSRESP:** (Vital signs) Respirations (per minute)
- **VSTEMP:** (Vital signs) Temperature
- **VSWEIGHT:** (Vital signs) Weight

# Bibliography


[1] Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's Disease Neuroimaging Initiative (ADNI)*, 2025. Accessed: January, 2025, from `https://adni.loni.usc.edu/`.

[2] Alzheimer's Disease Neuroimaging Initiative. *MRI Pre-processing*, n.d. Accessed: January, 2025 from `https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/mri-pre-processing/`.

[3] Adrian V. Dalca Bruce Fischl Malte Hoffmann Andrew Hoopes, Jocelyn S. Mora. *Synthstrip*, n.d. Accessed: February, 2025 from `https://surfer.nmr.mgh.harvard.edu/docs/synthstrip/`.

[4] Wiestler B. Navab N. Albarqouni S. Baur, C. Deep learning for unsupervised abnormal tissue detection in multi-contrast mri. *Medical Image Analysis*, 2020.

[5] Goh Joshua O. An Yang Kraut Michael A. O'Brien Richard J. Ferrucci Luigi Resnick Susan M. Beason-Held, Lori L. Changes in brain function occur years before the onset of cognitive impairment. *The Journal of Neuroscience*, Nov 2013.

[6] Seth Billiau. *From Scratch: Permutation Feature Importance for ML Interpretability*, 2021. Accessed: May, 2025 from urlhttps://towardsdatascience.com/from-scratch-permutation-feature-importance-for-ml-interpretability-b60f7d5d1fe9/.

[7] Kunik Mark E Schulz Paul Williams Susan P Singh Hardeep Bradford, Andrea. Missed, delayed diagnosis of dementia in primary care: Prevalence,

contributing factors. *Alzheimer Disease & Associated Disorders*, Oct–Dec 2009.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1), 2001.

[9] Jung Yun Jang Chandra A. Reynolds Nancy L. Pedersen Margaret Gatz Christopher R. Beam, Cody Kaneshiro. Differences between women, men in incidence rates of dementia, alzheimer's disease. *Journal of Alzheimer's Disease*, 2018. Accessed: 2025-04-22.

[10] Benedikt Wiestler Daniel Rueckert Julia A. Schnabel Andrew P. King Cosmin I. Bercea, Esther Puyol-Antón. Bias in unsupervised anomaly detection in brain mri. *arXiv preprint arXiv:2308.13861*, 2023.

[11] Matthew Dicicco. *Inertia For ML Applications*, 2022. Accessed: May, 2025 from `https://medium.com/@matthew.dicicco38/inertia-for-ml-applications-8c38de2d10d7`.

[12] Wei Yan Farnaz Nazari. Convolutional versus dense neural networks: Comparing the two neural networks' performance in predicting building operational energy use based on the building shape. *Building Simulation 2021 Conference Proceedings*, 2021.

[13] Frederik Rogge Julia Krüger Roland Opfer Alexander Schlaefer Finn Behrendt, Marcel Bengs. Unsupervised anomaly detection in 3d brain mri using deep learning with impured training data. *arXiv preprint arXiv:2204.05778*, 2022. Accepted for publication at the ISBI22 conference.

[14] GeeksforGeeks. *From Scratch: Permutation Feature Importance for ML Interpretability*, 2024. Accessed: May, 2025 from urlhttps://www.geeksforgeeks.org/ml-spectral-clustering/.

[15] GeeksforGeeks. *Mean Squared Error*, 2025. Accessed: May, 2025 from `https://www.geeksforgeeks.org/mean-squared-error/`.

[16] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. *arXiv preprint arXiv:1608.04667*, 2016. Accessed: 2025-04-30.

[17] Fujita Shohei Ohno Yoshiharu Aoki Shigeki Hagiwara, Akifumi. Variability, standardization of quantitative imaging: Monoparametric to multiparametric quantification, radiomics„ artificial intelligence. *Investigative Radiology*, Sep 2020.

[18] others Han, Y. Generative adversarial network for dementia detection using mri slice reconstruction. In *IEEE International Symposium on Biomedical Imaging*, pages 132–135, 2021.

[19] Matthey Loic Pal Arka Burgess Christopher Glorot Xavier Botvinick Matthew Mohamed Shakir Lerchner Alexander Higgins, Irina. beta-vae: Learning basic visual concepts with a constrained variational framework. *OpenReview*, 2017.

[20] Genichi Sugihara Jun Miyata Toshiya Murai Hidehiko Takahashi Manabu Honda Akitoyo Hishimoto Yuichi Yamashita Hiroyuki Yamaguchi, Yuki Hashimoto. Three-dimensional convolutional autoencoder extracts features of structural brain images with a 'diagnostic label-free' approach: Application to schizophrenia datasets. *Frontiers in Neuroscience*, 2021.

[21] Jorge Cadima Ian T Jolliffe. Principal component analysis: a review, recent developments. *The Royal Society*, 2016.

[22] Kaur Harleen Kaur, Manpreet. *Rating Scale of Silhouette Score*, 2024. Accessed: April 22, 2025 from `https://www.researchgate.net/figure/Rating-Scale-of-Silhouette-Score_tbl2_375082562`.

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

[24] Pär Kragsterman. The ultimate guide to preprocessing medical images: Techniques, tools„ best practices for enhanced diagnosis. *Collective Minds*, 2024.

[25] Hou Xianjun Lv, Jian. Alcoholism detection using hmi, predator-prey adaptive-inertia chaotic particle swarm optimization algorithm. *Computers in Biology, Medicine*, 2018.

[26] Xiaoran Chen Ender Konukoglu Shadi Albarqouni Matthäus Heer, Janis Postels. The ood blind spot of unsupervised anomaly detection. *OpenReview*, 2023.

[27] Healy John Melville James McInnes, Leland. Umap: Uniform manifold approximation, projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[28] Umberto Michelucci. An introduction to autoencoders. `https://arxiv.org/abs/2201.03898`, January 2022. Lecture notes; introductory paper.

[29] NI. *Structural Similarity Index*, 2023. Accessed: May, 2025 from `https://www.ni.com/docs/en-US/bundle/ni-vision-concepts-help/page/structural_similarity_index.html`.

[30] Nvidia. *K-Means Clustering Algorithm*, n.d. Accessed: May, 2025 from `https://www.nvidia.com/en-us/glossary/k-means/`.

[31] Open Genus. *Residual Connections in DL*, n.d. Accessed: May, 2025 from `https://iq.opengenus.org/residual-connections/`.

[32] Oxford Centre for Functional MRI of the Brain (FMRIB). *FLIRT module of FSL*, n.d. Accessed: January, 2025 from `https://fsl.fmrib.ox.ac.uk/fsl/docs/#/`.

[33] Mishra A. Patel, M. Pathological brain detection using hu moments, support vector machine. *International Journal of Computer Applications*, 2015.

[34] Duy Phuong. *Variational AutoEncoder*, 2018. Accessed: January, 2025 from `https://github.com/duyphuongcri/Variational-AutoEncoder`.

[35] Mechelli A. Sato J. R. Pinaya, W. H. L. Unsupervised deep learning reveals structural variations in neuropsychiatric disorders. *NeuroImage: Clinical*, 2021.

[36] Mechelli Andrea Sato João R. Pinaya, Walter H. L. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human Brain Mapping*, Feb 2019.

[37] Willens Victoria Prather Christina Moghtaderi Ali Chen Yi Gianattasio Kan Z. Grodstein Francine Shah Raj C. James Bryan D. Power, Melinda C. Risks, benefits of clinical diagnosis around the time of dementia onset. *Gerontology & Geriatric Medicine*, 2023.

[38] Davis D. N. Rahman, M. M. Alcoholism detection using data augmentation, convolutional neural networks, stochastic pooling. *Medical Imaging 2017: Image Processing*, 2017.

[39] Prince Asiamah Oluwatoyin Jolaoso Saroj Kumar, Ugochukwu Esiowu. Enhancing image classification with augmentation: Data augmentation techniques for improved image classification. *arXiv preprint arXiv:2502.18691*, 2025.

[40] Yannick Stephan Angelina R. Sutin Antonio Terracciano Selin Karakose, Martina Luchetti. Marital status, risk of dementia over 18 years: Surprising findings from the national alzheimer's coordinating center. *Alzheimer's & Dementia*, 2025. Accessed: 2025-04-22.

[41] Yuxiu Sui Shuai Liu Su-Jing Wang Yu-Dong Zhang Shui-Hua Wang, Yi-Ding Lv. Alcoholism detection by data augmentation, convolutional neural network with stochastic pooling. *Journal of Medical Systems*, 2018.

[42] Max Dünnwald Pavan Tummala Shubham Kumar Agrawal-Aishwarya Jauhari Aman Kalra Steffen Oeltze-Jafra Oliver Speck Andreas Nürnberger Soumick Chatterjee, Alessandro Sciarra. Strega: Unsupervised anomaly detection in brain mris using a compact context-encoding variational autoencoder. *arXiv preprint arXiv:2201.13271*, 2022.

[43] StackExchange Community. What is posterior collapse phenomenon?, n.d. Accessed: January, 2025 from `https://datascience.stackexchange.com/questions/48962/what-is-posterior-collapse-phenomenon`.

[44] The OECD Artificial Intelligence Policy Observatory. *Adjusted Rand Index (ARI)*, n.d. Accessed: May, 2025 from `https://oecd.ai/en/catalogue/metrics/adjusted-rand-index-ari`.

[45] Hinton G.E van der Maaten, L.J.P. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[46] others Veronica, J. Spatial autoencoders, variational models on diffusion mri for early parkinson's disease detection. *Frontiers in Neuroscience*, 2022.

[47] Blei David Cunningham John P Wang, Yixin. Posterior collapse, latent variable non-identifiability. In Y. Dauphin P.S. Liang J. Wortman Vaughan M. Ranzato, A. Beygelzimer, editor, *Advances in Neural Information Processing Systems*, volume 34, pages 5443–5455. Curran Associates, Inc., 2021.

[48] Jennifer L. Whitwell. Progression of atrophy in alzheimer's disease and related disorders. *Neurotoxicity Research*, 18(3-4), 2010.

[49] World Health Organization. *Dementia*, 2025. Accessed: April, 2025 from `https://www.who.int/news-room/fact-sheets/detail/dementia`.

University of Stavanger

4036 Stavanger
Tel: +47 51 83 10 00
E-mail: post@uis.no
www.uis.no

Cover Photo: Hein Meling