Group Project Report

Medical Imaging: Comparing Diseases

Iyanuloluwa Aketepe, Kyna Crumley, Kangou Diome, and Sandrine Soucy

Group Name: EMANON

SDS 3386: Data Science lab

Dr. Tanya Schmah

December 15, 2023

Table of Contents

[1 – Introduction 3](#_Toc153628480)

[2 – Datasets and Models 4](#_Toc153628481)

[2.1 – Models 4](#_Toc153628482)

[2.2 – Datasets 5](#_Toc153628483)

[3 – Analysis of the Autism dataset 7](#_Toc153628484)

[3.1 – Data Exploration 7](#_Toc153628485)

[3.2 – Dimensionality Reduction Techniques 8](#_Toc153628486)

[3.3 – Régression Logistique 9](#_Toc153628487)

[3.4 – K-Nearest Neighbours (KNN) 11](#_Toc153628488)

[3.5 – Support Vector Machine 11](#_Toc153628489)

[3.6 – Summary of the data analysis of the autism dataset 11](#_Toc153628490)

[4 – Analysis of the Alzheimer's Disease dataset 12](#_Toc153628491)

[4.1 – Data Exploration 12](#_Toc153628492)

[4.2 – Hypothesis Testing 13](#_Toc153628493)

[4.3 – Dimensionality Reduction Techniques 14](#_Toc153628494)

[4.4 – Classification from the studies model 15](#_Toc153628495)

[4.5 – Régression Logistique 16](#_Toc153628496)

[4.6 – K-Nearest Neighbours 18](#_Toc153628497)

[4.7 – Support Vector Machine 18](#_Toc153628498)

[4.8 – Linear Discriminant Analysis (LDA) 18](#_Toc153628499)

[4.10 – Summary of the different models 19](#_Toc153628500)

[5 – Analysis of the Preprocessed Alzheimer’s Disease dataset 19](#_Toc153628501)

[5.1 – Classification from the study’s model 19](#_Toc153628502)

[6 – Conclusion 20](#_Toc153628503)

[7 – Contributions 20](#_Toc153628504)

[References 21](#_Toc153628505)

# 1 – Introduction

The two questions that we addressed in this project were: can Convolutional Neural Network (CNN) models trained exclusively on Alzheimer's disease be generalized to new data or different diseases? And, which features extracted from structural MRI scans are most pivotal in distinguishing between diseased and non-diseased individuals?

Our methodology involves processing NIfTI files of structural MRI images using published Alzheimer’s disease detection models that output features. With an exploratory data analysis, we examined feature correlations and distributions, hypothesis testing, and utilized dimensionality reduction techniques. We also assess the effectiveness of various predictive models such as logistic regression, KNN, SVM, LDA, and classifier provided with the study.

We have used this methodology on three datasets: one comprising of 3 different degenerative diseases, one with autism patients and controls, and the other with Alzheimer's patients, mild cognitive impairment patients, and healthy controls.

By answering these questions, our goal is to find out if the model can be applied to different diseases and figure out the important features that help differentiate between disease states and healthy state.

# 2 – Datasets and Models

## 2.1 – Models

We planned to use two published models that were created to predict Alzheimer's disease in MRIs.

|  |
| --- |
| *A collage of diagrams and graphs  Description automatically generated* |
| *Figure 1. Overview of the deep learning framework and performance of the Generalizable deep learning model for early Alzheimer’s disease detection from structural MRIs* |

### 2.1.1 – Generalizable deep learning model for early Alzheimer’s disease detection from structural MRIs

Our first model was the 3D CNN published by Liu et al. in the Nature journal in 2022. This model aims to predict mild cognitive impairment and Alzheimer’s disease in individuals from their structural MRI. It consists of four convolutional blocks, each containing a convolutional layer, an instance normalization, a rectified linear unit (ReLU), and a 3D max pooling. More information about the convolutional layer can be found in Figure 1. For our analysis, we used a pretrained version of the model, which was trained on extensively preprocessed structural images from ADNI. This preprocessing included non-linear registration of the images to a DARTEL template computed from the training data, and then normalization to the Montreal Neurological Institute space using a Unified Segmentation procedure with Clinica. (Liu et al., 2020; Liu et al., 2022) This model achieved an AUC of 85.12 when discriminating between cognitively normal and Alzheimer’s disease/ Mild Cognitive Impairment patients from their preprocessed structural MRI images, according to Liu et al. The output of the model is a vector of length 1024, which we used to apply our own machine learning models, such as logistic regression, k-nearest neighbours (KNN) and support vector machine (SVM). These models will be discussed in detail in sections 3 and 4 for our data analysis.

### 2.1.2 – ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans

Another data model that would have helped us in our initial Alzheimer’s research idea was ADD-Net. ADD-Net is a data model published publicly by Fareed et al. on the IEEE Access journal in 2022. The model focuses on using machine learning to predict earlier forms of Alzheimer’s disease by training itself on a variety of Alzheimer’s 2-dimensional MRI scans. The model architecture consists of four convolutional blocks each with a 2D convolutional layer, a rectified linear unit and 2D average pooling. The model is not pre-trained, so an ample dataset of MRI scan images is also provided. The images used to train the dataset are formatted as jpg images, with a total size of 6,400 file. These MRI scan files are also separated into four distinct classes (“MildDemented”, ModerateDemented”, “NonDemented”, “VeryMildDemented”) specified in the following table:

|  |  |
| --- | --- |
| Alzheimer’s Disease MRI scan classes | Number of Image files per class |
| MildDemented | 896 files (~14% of dataset) |
| ModerateDemented | 64 files (~1% of dataset) |
| NonDemented | 3200 files (~50% of dataset) |
| VeryMildDemented | 2240 files (~35% of dataset) |

After the model has been manually trained, it will take as input a set of test data and predict each test image as one of the four classes above. We were planning to compare the predictions of this model and our other Alzheimer’s model to see what the differences would be in terms of how they predict Alzheimer’s. We also planned on seeing how well they would do on other neurodegenerative diseases like Parkinson’s disease, Traumatic Brain Injury, etc though we encountered some challenges that changed our end project. One of the issues we had to deal with was understanding how to convert the 3D Nifti images of the first model into a 2D png images for the ADD-Net model. While we were able to make ground on the topic, the amount of time we needed to pursue the idea was not in line with course expectations. Another problem was doing any data transformation with the 3D images in general. The files required an extremely high amount of computing power which resulted in greater time loss.

## 2.2 – Datasets

### 2.2.1 – Alzheimer, Parkinson, and Multiple Sclerosis

Our first dataset was a mix of five different datasets. The first one, *Brain MRI Dataset of Multiple Sclerosis with Consensus Manual Lesion Segmentation and Patient Meta Information,* from which we only used the structural (T1-weighted) MRI images (n = 60). (M Muslim, 2022) The second dataset, was a combined dataset of healthy control subjects, subjects with bipolar disorder, and subjects with Alzheimer’s disease, from which we used T1 images of the healthy control subjects (n = 20) and Alzheimer’s disease patients (n = 26). (Besga et al., 2020) The third data set we used all T-1weighted images of healthy aging subjects (n = 25) and Parkinson’s patients (n = 21). (Day et al., 2023) From the fourth (n = 40), and the fifth (n = 58) datasets we used all T1-weighted images from all subjects, all of which were diagnosed with Parkinson’s disease. (Abbass et al., 2023; Wylie et al., 2023)

While all voxel intensities are normalized and scaled before being evaluated by the model, our dataset contained raw images of the Alzheimer’s disease patients and healthy control subjects from the second dataset, and preprocessed images from the first dataset of Multiple Sclerosis patients. This preprocessing included skull-stripping for the anonymization of the subject and ease of analysis, which leaves only brain matter in the image. The difference between the raw images and the preprocessed images was apparent when applying dimensionality reduction techniques to visualize the two groups. Figure 2 shows that the convolutional neural network model for early detection of Alzheimer’s disease can clearly discriminate between raw images and preprocessed images, as t-SNE allows for clear separation of the two groups. (Liu et al., 2022)

A graph with many dots

Description automatically generated

*Figure 2. t-SNE visualization of features extracted by the Convolutional Neural Network (CNN) model from MRI images of patients in the Multiple Sclerosis (MS) dataset, Alzheimer's Disease (AD) patients, and healthy controls (HC). The plot demonstrates two clearly distinct groups, highlighting the ability of the CNN model to capture the differences in image features among images from the preprocessed MS dataset and the raw images from the AD and healthy control dataset. This visualization suggests that the CNN features effectively discriminate between the preprocessed images and the raw images.*

Image preprocessing, including steps like skull-stripping and non-linear registration, plays a pivotal role in any machine learning-based analysis of MRI images. These preprocessing steps are essential for ensuring data quality, consistency, and alignment across subjects. Skull-stripping removes extraneous non-brain tissue, enhancing the model's ability to focus on relevant brain structures. Non-linear registration aligns images to a common template, enabling meaningful inter-subject comparisons. These processes not only improve the accuracy and robustness of machine learning models but also enable the extraction of biologically meaningful features, leading to more reliable and interpretable results in applications such as disease diagnosis. In essence, image preprocessing is the foundation upon which accurate and informative MRI-based machine learning analyses are built.

Based on time and ability constraints, preprocessing the images from the datasets containing raw images was not feasible. In order to find any interesting results, the project required shifting to using only images from one study.

We also briefly used a subset of the preprocessed images from the Alzheimer’s disease vs. Healthy Control vs. Bipolar Disorder, which included 54 T1-weighted preprocessed images of the Alzheimer’s disease patients (n = 35) and healthy controls (n = 19) in order to verify whether our analysis was hindered by the lack of preprocessing. These images were preprocessed by registering the T1-weighted volumes to the Montreal Neurological Institute template using FSL non-linear registration tools (DARTEL). (Besga et al., 2020) We utilized the same methods to extract features from the images.

### 2.2.2 – Autism Brain Imaging Data Exchange (ABIDE)

Our second dataset comes from the Autism Brain Imaging Exchange, which was published in 2013 by Di Martino et al. This dataset is publicly available online and includes a total of 1112 individuals for which the scans were provided by multiple universities in the United States. Of these individuals, 539 have been diagnosed with autism spectrum disorders (ASD) while the remaining 573 are typical controls. The dataset includes both resting-state functional magnetic resonance imaging (R-fMRI) and structural MRI data. However, for our analysis, we focused only on the structural MRI data. The phenotypic data was available for all 1112 individuals and included sex, age at scan, IQ, scores of tests used for diagnosing autism such as Autism Diagnosis Observation Schedule (ADOS), Social Responsiveness Scale (SRS) and Autism Diagnostic Interview (ADI). For our analysis, our columns of interest were the binary diagnosis which indicates if the individual is diagnosed with an ASD or is a control, the age at scan and the sex. Also, we were interested on evaluating the relationship between the MRI and the different scores from the diagnostic information.

### 2.2.3 – Alzheimer's Disease Neuroimaging Initiative (ADNI)

We used a subset of structural MRI images from the ADNI1\* phase, totaling 558 MRI scans collected from 294 subjects. It's worth noting that some subjects contributed multiple images at different time-points during the study. Each 3D MRI image was initially stored as a collection of approximately 120 to 180 DICOM files, representing individual slices composing the final image. The conversion of these DICOM image sets into NifTi files for evaluation with our pre-trained convolutional neural network presented challenges. The primary challenge arose from inconsistencies encountered in an application used for NifTi image viewing, resulting in the perceived introduction of artifacts not being present in the images. Consequently, extensive troubleshooting efforts were undertaken, which included the use of various software packages for .dcm to .nii conversion, as well as attempts to develop a custom function to rectify the unexpected artifact. These efforts aimed to ensure the accuracy and reliability of our subsequent data analysis.

Due to time and ability constraints, the images were not preprocessed before model evaluation, however this step is crucial to enable the extraction of biologically meaningful features and would be important for any future work.

Data was extracted from a clinical dataset contained in a comma-separated-values file to match estimated age of the subject, calculated using a known age at a known date with the date the MRI was performed, with each image in order to pass both the subject’s age on the date of the scan and the image as a NiFti file to the model. The 1024 features extracted from each image by the model, as well as the subject’s age, ID, and the date of the image were then consolidated into one dataset. Then the subject’s diagnosis (Alzheimer’s Disease, Mild Cognitive Impairment, or Cognitively Normal) at the closest recorded date in the clinical data file to the date of the MRI was added, as well as if that subject ID was recorded as being used for training, validation or testing as published by Liu et al. of the model. The final dataset consists of 523 rows of observations, which is reduced from the number of images due to errors in

image conversion and model evaluation, and 1031 columns.

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

# 3 – Analysis of the Autism dataset

## 3.1 – Data Exploration

We first started with the exploratory data analysis. Although it is a simple task, it is much more difficult with 1024 continuous columns.

|  |
| --- |
| *A graph with red squares  Description automatically generated* |
| *Figure 3. Correlation matrix of the first 50 features of the autism dataset* |

### 3.1.1 – Correlation

Our first analysis focuses on the correlation between the features. Since our dataset contains a large number of features (1024), which is nearly equal to the number of rows, dimensionality reduction techniques are crucial for model selection later on. As seen in correlation matrix in Figure 3, the pairwise correlation coefficients between the features are very low. Indeed, the distribution of the pairwise correlation coefficients follow a normal distribution with a mean very close to 0, a standard deviation of 0.03, a minimum correlation coefficient of -0.136 between feature #331 and feature #417 and a maximum correlation coefficient of 0.140 between feature #633 and feature #737. Consequently, our dataset contains highly uncorrelated features that lack a strong linear relationship.

### 3.1.2 – Distributions of the features

As seen in Figure 4, the features follow a normal distribution centered around 0, ranging between -2 to 2. This is beneficial for statistical testing since some have the assumptions that the data is to be normally distributed.

|  |
| --- |
| *A group of graphs showing different sizes of data  Description automatically generated with medium confidence* |
| *Figure 4. Histogram of the first 10 features by diagnosis of the autism dataset* |

## 3.2 – Dimensionality Reduction Techniques

As mentioned above, our dataset contains 1024 features for 1112 rows making dimensionality reduction an important step before the analysis.

|  |  |
| --- | --- |
|  |  |
| *Figure 5. Scree plot of the first 50 principal components of the autism dataset* | *Figure 6. Line plot of the cumulative explained variance of the first 700 principal components of the autism dataset* |

### 3.2.1 – Principal Component Analysis (PCA)

Our first dimensionality reduction technique that we tried was principal components analysis (PCA). In the scree plot pictured in Figure 5, we can see that all the principal components explained 0.4% or less of the variance each. In total, as seen in Figure 6, it would take 200 principal components to explain 50% of the variance of the original dataset, 550 principal components for 90%, and 650 principal components for 95%. These results were to be expected, because of the theory behind PCA. PCA’s intention is to transforms a large number of correlated variables into a new set of uncorrelated variables which are the principal components which capture the maximum variance of the original data. In our dataset, as shown above, we have highly uncorrelated data to start with making PCA not a feasible dimensionality reduction technique for our dataset.

|  |  |
| --- | --- |
| *A diagram of a diagram with blue and orange dots  Description automatically generated* | *A diagram of a diagram with blue and orange dots  Description automatically generated* |
| *Figure 7. Scatterplot of the first 2 principal components* | *Figure 8. Scatterplot of the t-SNE components* |

### 3.2.2 – t-distributed Stochastic Neighbour Embedding (t-SNE)

Since our dataset had very low pairwise correlation, we decided to test with t-distributed Stochastic Neighbour Embedding (t-SNE) as our second dimensionality reduction technique. This technique can often capture non-linear relationships between the variables; however, it unfortunately did not work with our dataset and the results can be seen in Figure 8.

## 3.3 – Régression Logistique

La régression logistique est une technique statistique utilisée pour modéliser la probabilité d'une variable dépendante binaire (y pour notre dataset: 1 pour autisme et 0 pour sain) en fonction de variables indépendantes (x1 à x1024). Elle est largement utilisée en classification, notamment dans des domaines tels que l'apprentissage automatique, la statistique et l'épidémiologie. Voici quelques points clés concernant la régression logistique et son importance dans le contexte de notre projet.

### 3.1 Modélisation de la Probabilité

La régression logistique modélise la probabilité que la variable dépendante prenne la valeur 1 (ou une classe positive) en fonction des variables indépendantes. Ici, nos variables indépendantes sont les caractéristiques pour détecter une maladie (autisme) autre que l’alzheimer. Elle utilise la fonction logistique pour transformer une combinaison linéaire des variables indépendantes (x1 à x1024 qui ont été réduits grâce au PCA) en une probabilité.

### 3.2 Interprétabilité

Les coefficients de régression dans la régression logistique fournissent des informations sur la force et la direction de la relation entre les caractéristiques de la maladie d’alzheimer et la variable dépendante y (autisme). Cela permet d'interpréter comment chaque variable contribue à la prédiction de l’autisme.

### 3.3 Évaluation de la Performance

Les mesures de performance telles que la précision, le rappel, la courbe ROC (Receiver Operating Characteristic), et l'aire sous la courbe (AUC) sont souvent utilisées pour évaluer la performance d'un modèle de régression logistique.

### 3.4 Observations

La précision du modèle est de 0.48, cela signifie que le modèle parvient à prédire la classe correcte pour environ 48% des échantillons dans l'ensemble de test mais pas plus. Cependant, cela ne fournit pas nécessairement une image complète de la performance du modèle ; ce qui impacte sur nos attentes négativement et il est souvent utile d'examiner d'autres métriques.

Voici quelques observations à souligner :

Précision du modèle après PCA: 0.47058823529411764

|  |
| --- |
| *A black and white screen with numbers  Description automatically generated* |
| *Figure 9* |

Performance Modérée : La précision de 0.48 indique une performance modérée. Cela pourrait être dû à plusieurs raisons, telles que des caractéristiques insuffisamment informatives comme nos images utilisés qui diffèrent, une complexité de modèle inappropriée (on a rencontré un problème avec l’application de PCA sur notre dataset), et aussi des problèmes dans la qualité des données.

Évaluation avec d'autres Métriques : En plus de la précision, il est recommandé d'examiner d'autres métriques telles que le rappel, la F1-score, et la matrice de confusion. Ces métriques fournissent une image plus complète de la performance, surtout dans le cas de classes déséquilibrées (les images ici).

|  |
| --- |
| *A blue squares with white text  Description automatically generated* |
| *Figure 10* |

Les résultats de notre modèle après PCA montrent les éléments suivants :

Rappel (Recall) : Le modèle a un rappel d'environ 48,61%. Cela signifie que sur l'ensemble des exemples positifs réels, le modèle en a identifié correctement environ 48,61%. Un rappel relativement faible indique que le modèle a du mal à capturer toutes les instances positives c’est à dire les ressemblances entre les sujets.

F1-score : Un F1-score de 48.72% indique une performance relativement équilibrée entre la capacité du modèle à classifier correctement les exemples positifs (cerveaux malades) et les exemples négatifs (cerveaux sains).

NB: Dans le contexte médical, les conséquences des faux positifs (prédire à tort un cerveau malade) et des faux négatifs (ne pas détecter un cerveau malade) peuvent être différentes. Certains scénarios peuvent nécessiter une focalisation particulière sur la précision ou le rappel, en fonction des exigences de la tâche.

Matrice de Confusion : La matrice de confusion montre le nombre de vrais positifs (229), de vrais négatifs (210), de faux positifs (222) et de faux négatifs (220). Cela offre une vue détaillée des performances du modèle.

Interprétation en bref :

Le modèle a tendance à classer de manière incorrecte un nombre significatif d'exemples positifs.

La performance globale, mesurée par le F1-score, est modérée.

Cette interprétation suggère que le modèle actuel a des limitations dans sa capacité à bien généraliser et à identifier correctement les autistes. Des ajustements et une exploration plus approfondie sont recommandés pour améliorer les performances du modèle.

Importance des Caractéristiques : La réduction de dimensionnalité par PCA peut entraîner une perte d'information. Il pourrait être intéressant de vérifier l'importance des caractéristiques pour s'assurer que les caractéristiques conservées sont réellement informatives.

En résumé, bien que la précision soit une mesure importante, il est essentiel de l'évaluer en conjonction avec d'autres métriques et d'effectuer une analyse plus approfondie pour comprendre les raisons sous-jacentes de la performance du modèle.

## 3.4 – K-Nearest Neighbours (KNN)

K-Nearest Neighbours was one of the models we had chosen for our predictions. It relies on memorizing the characteristics of training dataset datapoints and uses a user-defined number (K) to pick the closest training datapoints to a given test datapoint and use the most common outcome to assign a prediction.

|  |
| --- |
| *A graph with a line  Description automatically generated* |
| *Figure 11* |

Initially, the KNN was only set to be used for one K value. However, we thought it would be more interesting to show how changes in K would affect the accuracy of our model. So, we ran multiple KNN-based predictions with a K range of [3, 15]. Unfortunately, our results did not showcase anything drastically different from the first run-through. Overall, the accuracy stays relatively the same, with the minimum accuracy being ~45% (K = 4) and the maximum accuracy being ~52% (K = 9). Overall, the mean accuracy was ~48%.

|  |
| --- |
| A screenshot of a computer  Description automatically generated |
| *Figure 13* |

|  |
| --- |
| *A screenshot of a computer screen  Description automatically generated* |
| *Figure 12* |

## 3.5 – Support Vector Machine

Support Vector Machines was another model we used for making predictions on our data. It focuses on trying to take all the characteristics of the training data and use that to make the most decisive split between the datapoints. This split is what is used to assign class predictions to datapoints.

Unfortunately, results for the SVM were underwhelming, with an accuracy of ~45%.

## 3.6 – Summary of the data analysis of the autism dataset

In conclusion, our results for autism dataset were very poor. We can expect that those results are since the model use is for Alzheimer’s disease, was trained on Alzheimer’s disease and there is significant difference between the brain scan of an individual with an autism spectrum disorder and an individual with Alzheimer’s disease therefore the model is not trained to recognize the aspects of autism in a brain scan.

|  |
| --- |
| *A diagram of a number of numbers  Description automatically generated with medium confidence* |
| *Figure 14. Correlation Matrix of the first 50 features of the Alzheimer dataset* |

# 4 – Analysis of the Alzheimer's Disease dataset

## 4.1 – Data Exploration

### 4.1.1 – Correlation

We started again with correlation analysis. In Figure 14, we can see that the features have stronger positive and negative correlations between the features than the autism dataset. The pairwise correlation coefficients follow once again a normal distribution centered around 0 there is more variation with the standard deviation being 0.17 with minimum correlation coefficient of -0.705 between feature #39 and feature #964, and maximum correlation coefficient of 0.815 between feature #39 and feature #79.

|  |  |
| --- | --- |
|  |  |
| *Figure 15. Histogram of the first 6 features of Alzheimer’s dataset* | *Figure 16. Histogram of the first 6 features scaled of Alzheimer’s dataset* |

### 4.1.2 – Distributions of the features

As seen in Figure 15, the features follow a somewhat normal distribution centered around 0, ranging between -2 and 2. Scaling the features data to have a mean of 0 and a standard deviation of 1, we can see better normal distributions centered around 0, ranging between -4 and 4. In fact, in Figure 17, we have the quantile-quantile (Q-Q) plots of some of the first 6 features and it is clear that they follow a normal distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| *Figure 17. Quantile-Quantile plot of the first 5 features scaled of the Alzheimer’s dataset* | | | | |

|  |
| --- |
| *A group of graphs with different colored lines  Description automatically generated with medium confidence* |
| *Figure 18. Boxplot by diagnosis of the features with p-value of the ANOVA test under 0.001* |

## 4.2 – Hypothesis Testing

|  |
| --- |
| *A group of boxes with numbers and symbols  Description automatically generated with medium confidence* |
| *Figure 19. Boxplot by diagnosis of the features with p-value of the ANOVA test under 0.001* |

We then proceeded to see if there are significant differences between the 3 categories of diagnosis for each feature. Since we have more than 2 categories, we final choice was between ANOVA and Kruskal- Wallis. We chose ANOVA since with the Q-Q plot of Figure 17, we have the normality assumption. With setting a significance value at 0.01, we have 18 features that have p-values from the ANOVA test under that significance level. Shown in figure 18 are the boxplot of the features with a p-value under 0.005 indicating that one of the categories is significantly different than the others.

We then tried ANOVA tests for each feature with only two of the categories: healthy control and Alzheimer’s disease individuals. When setting a significance value at 0.01, we have 14 features that have a p-value from the ANOVA test under that significance level. In figure 19, we have the boxplot of the features with a p-value under 0.005 indicating that the mean of the Alzheimer’s disease patients and the healthy control are significantly different for those features.

|  |  |
| --- | --- |
| *A graph with a green line  Description automatically generated* | *A graph with a green line  Description automatically generated* |
| *Figure 20. Scree plot of the first 30 principal components of the Alzheimer’s dataset* | *Figure 21. Line plot of the cumulative explained variance of the first 200 principal components of the Alzheimer’s dataset* |

## 4.3 – Dimensionality Reduction Techniques

### 4.3.1 – Principal Components Analysis (PCA)

We then tried dimensionality reduction techniques starting with PCA. As it can be seen in the scree plot in Figure 20, the first principal component explains ~12% of the variance of the original followed by ~6% for the second principal component, ~5% for the third principal component, and less than 5% for the others. In the cumulative explained variance plot in figure 21, we can see that it would take 130 principal components to explain 90% of the variance and 170 principal components to explain 95% of the variance. As it can be seen in figure 22 and figure 23, the first principal components do not create clusters, and don’t separate one diagnosis to the others.

|  |  |
| --- | --- |
|  |  |
| *Figure 22. Scatterplot of the first and second principal components* | *Figure 23. Scatterplot of the third and fourth principal components* |

### 4.3.2 – t-distributed Stochastic Neighbour Embedding (t-SNE)

Our second method of dimensionality reduction was t-SNE with the three diagnoses. We had expectation for the result of this method since the published model has their t-SNE projection in their paper (Liu, 2022). However, our results, in Figure 24, were drastically different and did not create similar clusters like their results, in Figure 25, using the default parameter of t-SNE from sklearn.

|  |  |  |
| --- | --- | --- |
| *A graph with different colored dots  Description automatically generated* |  | *A graph with blue and orange dots  Description automatically generated* |
| *Figure 24. Our t-SNE projection* | *Figure 25. The t-SNE project in the published study (Liu, 2022)* | *Figure 26. The t-SNE projection for Alzheimer and healthy patients only* |

After, we tried t-SNE without the mild cognitive impairment diagnosis. We can see in figure 26 that there are no specific clusters, however it seems like there could be a further separation between the Alzheimer’s disease individuals and the healthy control if we look further into the differences.

|  |
| --- |
| *A graph with blue and orange dots  Description automatically generated* |
| *Figure 27. Scatterplot of the two UMAP components by diagnosis.* |

### 4.3.3 – Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is another type of dimensionality reduction technique that allows for non-linear relationships. Continuing with only the Alzheimer’s disease and healthy individuals, we proceeded with a UMAP. Testing multiple distance metrics, number of neighbours, minimum distances, and combination of features, we were able to have a decent looking graph that has mostly Alzheimer’s disease individuals at the top and healthy control on the bottom which is shown in Figure 27. The features used are the features that had a p-value under 0.01 for the ANOVA test presented in Figure 19.

## 4.4 – Classification from the study'model

The implementation of the provided classifiers with the CNN model was a comprehensive process, which involved meticulous hyperparameter tuning through grid search to ascertain the most effective hyperparameters for learning to classify between patients with Alzheimer’s disease and cognitively normal subjects. The model is initially trained on a dataset, followed by a validation phase for fine-tuning hyperparameters, and ultimately, it is tested on an unseen dataset to evaluate its efficacy. We employ the Area Under the Curve (AUC) metric for performance evaluation, a robust measure in classification tasks for balancing true positive and false positive rates. Additionally, the incorporation of cross-entropy as the loss function is a strategic choice for handling probabilistic outcomes in classification tasks. The Stochastic Gradient Descent (SGD) optimizer is utilized for its proven efficiency and simplicity, further enhancing the model's learning capability.

### 4.4.1 – Linear Classifier from the Study’s Model

The Linear Classifier is a PyTorch neural network model designed for classification of linearly separable data. It consists of a sequential container that initially flattens the input, then applies a linear transformation from the input dimension to the number of hidden layers, followed by another linear layer from the hidden layers to the output dimension. The initialization initializes the weights of the linear layers with a normal distribution and biases to zero, which helps for faster convergence in training. This model is quite simple, focusing on linear transformations, and emphasizing stable and efficient training.

Despite meticulous hyperparameter tuning, and also using the hyperparameters used by the original study, and employing robust measures like Area Under the Curve (AUC) for performance evaluation, the model achieved an AUC of 0.60 with a wide 95% confidence interval ranging from 0.40 to 0.80. An AUC score of 0.60, while slightly better than a random guess, is not particularly high, indicating only a low ability of the model to distinguish between subjects with Alzheimer’s disease and cognitively normal subjects from their structural MRIs. The broad confidence interval further implies significant uncertainty in the model’s performance. The lower bound of the confidence interval at 0.40 suggests the possibility that the model’s performance is no better than flipping a coin.

Several factors could be contributing to these relatively poor results. Firstly, unlike the original study, our approach did not involve preprocessing of MRI images. This is especially important considering that the convolutional neural network was trained on extensively preprocessed MRI images. Secondly, we utilized a smaller subset of the original dataset. The reduced size of the dataset might have affected the classifier’s ability to learn effectively, potentially leading to overfitting or underfitting.

Statistically, the model cannot be conclusively deemed effective for the task of distinguishing between the two classes from unprocessed MRI images, given the large confidence interval for the AUC score. The results suggest that further improvements are necessary, possibly through more extensive data preprocessing, training the feature extraction model on raw images, or using a larger dataset.

### 4.4.2 – Adversarial Classifier from the Study’s Model

Adversarial classification is a machine learning technique that involves training a model to make it robust against adversarial examples – inputs that would normally confuse or deceive the model. In the context of medical imaging, adversarial classification can be intriguing because it potentially improves the model’s resilience against variations and noise in raw images. We hypothesized that this resilience might be beneficial for our context, where image preprocessing was not done.

The Adversarial Classifier is a PyTorch neural network model designed for classification. The model is initialized with input dimension, hidden layer size, and output dimension, and it has a relatively simple structure, comprising a single linear layer from the input dimension to the output dimension. The forward pass involved normalizing the output of the linear layer using an L2 normalization. The normalization step ensures that the output vectors have a unit norm, which is beneficial for learning stability and performance in an adversarial training context. The architecture is straightforward without hidden layers or non-linear activations, which means that it was designed for linearly separable data.

In our attempt to replicate the original study’s work using a smaller sample size and without preprocessing the MRI images, the trained classifier achieved an AUC of 0.61, with a 95% confidence interval ranging from 0.42 to 0.79. While the AUC of 0.61 indicates the possibility that the model may perform slightly better than random guessing, the wide confidence interval (0.42 to 0.79) suggests an extremely high degree of uncertainty in the model’s effectiveness in discriminating between the two classes. The lower bound of the confidence interval raises concerns about the model’s consistency and reliability.

While the adversarial classification approach shows some potential, the current model requires further refinement and exploration before it can be considered effective for the detection of Alzheimer’s disease from structural MRIs. For future iterations, several improvements can be considered, including an extensive preprocessing pipeline, larger sample size, or a more complex model architecture in order to better capture the nuances of the data.

## 4.5 – Régression Logistique

Pour cette partie on a procédé de 4 manières pour trouver le meilleur modèle avec différentes types de régressions.

Les modèles 1 à 3 n’ont rien donné et même un changement du code n’a pas aidé. On a préféré mettre les résultat pour donné une vue globale de nos régressions.

* Modéle 1 une régression logistique simple: 10%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rapport de classification | | | | |
|  | precision | recall | f1-score | support |
| 1 | 0.00 | 0.00 | 0.00 | 37 |
| 2 | 0.10 | 1.00 | 0.19 | 5 |
| 3 | 0.00 | 0.00 | 0.00 | 6 |
| accuracy |  |  | 0.10 | 48 |
| macro avg | 0.03 | 0.33 | 0.06 | 48 |
| weigthed avg | 0.01 | 0.10 | 0.02 | 48 |

Précision du modèle: 0.10

|  |  |  |
| --- | --- | --- |
| Matrice de confusion | | |
| 0 | 37 | 0 |
| 0 | 5 | 0 |
| 0 | 6 | 0 |

* Modéle 2 une Lasso Logistique Régression: 10%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rapport de classification | | | | |
|  | precision | recall | f1-score | support |
| 1 | 0.00 | 0.00 | 0.00 | 37 |
| 2 | 0.10 | 1.00 | 0.19 | 5 |
| 3 | 0.00 | 0.00 | 0.00 | 6 |
| accuracy |  |  | 0.10 | 48 |
| macro avg | 0.03 | 0.33 | 0.06 | 48 |
| weigthed avg | 0.01 | 0.10 | 0.02 | 48 |

Results for Lasso Logistic Regression: Accuracy of the model: 0.10

|  |  |  |
| --- | --- | --- |
| Matrice de confusion | | |
| 0 | 37 | 0 |
| 0 | 5 | 0 |
| 0 | 6 | 0 |

* Modèle 3 Ridge Logistique Régression: 10%

Results for Ridge Logistic Regression: Accuracy of the model: 0.10

|  |  |  |
| --- | --- | --- |
| Matrice de confusion | | |
| 0 | 37 | 0 |
| 0 | 5 | 0 |
| 0 | 6 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rapport de classification | | | | |
|  | precision | recall | f1-score | support |
| 1 | 0.00 | 0.00 | 0.00 | 37 |
| 2 | 0.10 | 1.00 | 0.19 | 5 |
| 3 | 0.00 | 0.00 | 0.00 | 6 |
| accuracy |  |  | 0.10 | 48 |
| macro avg | 0.03 | 0.33 | 0.06 | 48 |
| weigthed avg | 0.01 | 0.10 | 0.02 | 48 |

|  |
| --- |
| *A graph with different colored bars  Description automatically generated* |
| *Figure 28* |

Par la suite, une meilleure solution pour nous a été de changer les caractéristiques prédits voir en enlever certains. On obtient alors notre 4em model.

Feature Combinations X = 'PC1', 'PC2'

* Modèle 4 : **65%**

Les 3 premières méthodes n'ont pas donné de résultat optimal et pratiquement ont fournies les mêmes résultats d'analyse. Par contre en effectuant une combinaison des features pour obtenir le modèle 4, on obtient 65% de précision (accuracy) ce qui est beaucoup meilleur pour nous.

Confusion Matrix (Matrice de Confusion):

|  |
| --- |
| *A blue square with white text  Description automatically generated* |
| *Figure 29* |

Les vrais positifs (TP) sont les cas où le modèle prédit correctement la classe 2.0 (12 cas).

Les vraies negatives (TN) sont les cas où le model prédit correctement les classes 1.0 et 3.0 (0+1 cas).

Les faux positifs (FP) sont les cas où le model prédit à tort la classe 2.0 (3 cas).

Les faux négatifs (FN) sont les cas où le model prédit à tort les classes 1.0 et 3.0 (1+2 cas).

Cette matrice fournit une vue détaillée des performances du model en termes de classifications correctes et incorrectes.

Classification Report (Rapport de Classification):

* Precision (Précision): Pour la classe 1.0, la précision est de 0.00 (pas de vrais positifs). Pour la classe 2.0, la précision est de 0.71 (12 vrais positifs). Pour la classe 3.0, la précision est de 0.50 (1 vrai positif). La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives.
* Recall (Rappel): Pour la classe 1.0, le rappel est de 0.00 (pas de vrais positifs). Pour la classe 2.0, le rappel est de 1.00 (12 vrais positifs). Pour la classe 3.0, le rappel est de 0.25 (1 vrai positif). Le rappel mesure la proportion de vrais positifs parmi toutes les observations réellement positives.
* F1-score (Score F1): C'est une mesure combinée de la précision et du rappel. Il est de 0.83 pour la classe 2.0, ce qui est généralement considéré comme un bon score.
* Support: Le nombre d'occurrences réelles de chaque classe dans l'ensemble de données de test.

En résumé, le modèle a une bonne précision et rappel pour la classe 2.0, mais des performances inférieures pour les autres classes. Ces résultats suggèrent que le modèle peut être plus performant pour la classe 2.0 par rapport aux autres classes.

|  |
| --- |
| *A graph with a line  Description automatically generated* |
| *Figure 30* |

## 4.6 – K-Nearest Neighbours

The KNN accuracy results we had received were slightly better than our autism data results, though both are still around the 50% accuracy mark. Overall, the accuracy stays relatively the same, with the minimum accuracy being ~48% (K = 8) and the maximum accuracy being ~60% (K = 4). Overall, the mean accuracy was ~48%.

## 4.7 – Support Vector Machine

Accuracy achieved was ~42%.

|  |  |
| --- | --- |
| A screenshot of a computer screen  Description automatically generated | *A screenshot of a computer  Description automatically generated* |
| *Figure 31.* | *Figure 32.* |

|  |  |  |
| --- | --- | --- |
| Confusion Matrix | | |
|  | AD | HC |
| AD | 12 | 6 |
| HC | 2 | 16 |

|  |  |  |
| --- | --- | --- |
|  | Precision | Recall |
| AD | 0.86 | 0.67 |
| HC | 0.73 | 0.89 |

## 4.8 – Linear Discriminant Analysis (LDA)

With the UMAP projection of Alzheimer’s disease individuals and healthy controls, a good model to try was Linear Discriminant Analysis (LDA) since it has a clear possible decision boundary. We fitted the LDA with a 70/30 split for training and test, and we got an accuracy of 77.8% for the test set.

However, this good result was not replicated on new data. As seen in figure 34, the clusters presented in Figure 33 are not the same. The accuracy of the model on the new data was 48.7% meaning that the original model is overfitting the training data.

|  |  |
| --- | --- |
| *A graph showing different colored circles  Description automatically generated with medium confidence* |  |
| *Figure 33. Scatterplot of the two UMAP components with LDA decision boundary on training and test set* | *Figure 34. Scatterplot of the two UMAP components with LDA decision boundary on new data* |

## 4.10 – Summary of the different models

The analysis of the Alzheimer's Disease dataset involved looking at things like feature relationships, distributions, and hypothesis testing. This helped us find some important differences between the different diagnostic categories. We also used some techniques to simplify the data like PCA, t-SNE, and UMAP, but it was hard to find clear clusters. When we tried different classification models, we found that the AUC scores were only around 0.60. At first, logistic regression models didn't work very well either, but when we used a combination of relevant features, we were able to improve the performance to 65%. The K-Nearest Neighbours and Support Vector Machine models had accuracies around 50% and 42%, respectively. Linear Discriminant Analysis had good accuracy on the test set (77.8%), but it didn't work as well with new data (48.7%). These findings show that classifying Alzheimer's disease is really complicated, so we need to keep refining and exploring our models to make them better.

# 5 – Analysis of the Preprocessed Alzheimer’s Disease dataset

## 5.1 – Classification from the study’s model

We implemented a comprehensive evaluation of the two neural network classifiers ‘LinearClassifierAlexNet’ and ‘AdversarialClassifier’ on the preprocessed dataset (n = 54), in order to gain more insight into whether our relatively poor classification results were affected by the lack of preprocessing of the ADNI images. This is because this dataset was preprocessed in a similar manner to the images on which the o CNN model was originally trained and evaluated.

Given the potential class imbalance (19 cognitively normal subjects : 35 Alzheimer's disease patients), class weights were computed and used during the training of the models to provide a more balanced learning process. To account for variability in model performance, bootstrapped confidence intervals for the AUC were computed, and then stratified k-fold cross-validation was applied to each model to validate its performance more robustly on different subsets of the preprocessed dataset. The bootstrapped confidence intervals (CI) for the Area Under the ROC Curve (AUC) reflect significant variability in model performance. For the linear classifier, the AUC was calculated to be 0.55, and the CI ranges from 0.5 to approximately 0.93, while the adversarial classifier the CI ranges from 0.41 to 0.85. The cross-validated confidence intervals yielded higher mean AUCs for both models, with the linear classifier achieving a mean AUC of approximately 0.76 (95% CI: 0.556 - 0.959) and the adversarial classifier achieving a mean AUC of approximately 0.70 (95% CI: 0.407 - 0.988).

Based on the analysis with the preprocessed images, the convolutional neural network model does show potential in discriminating between cognitively normal subjects and subjects with Alzheimer’s disease, especially since the lower bound 95% confidence interval of the linear classifier is significantly higher than 0.5. Further analysis would be required, as shown by the wide confidence intervals in both bootstrapped and cross-validated evaluations highlight the variability and uncertainty in model performance, likely exacerbated by the small sample size and class imbalance.

# 6 – Conclusion

In summary, our project set out to assess the generalizability of currently published models trained on Alzheimer's disease to different diseases and identify key features from MRI scans for disease classification. Despite encountering challenges with dataset availability, data preprocessing limitations, and time constraint, we still conducted a comprehensive analysis of both the autism dataset and the Alzheimer's disease datasets. The the analysis of the autism dataset yielded poor results, and we were unable to distinguish between cognitively normal subjects and subjects with autism from the features extracted by the model. This is most likely because the difference in brain structure between these two classes is subtle, and the model was unable to capture these subtle differences in the features, especially without image preprocessing. However, the Alzheimer’s disease analysis had some interesting results. We found stronger correlations in the features, some dimensionality reduction techniques could be tested further to see possible separation between the diagnosis, and some of the results of the classification model were better if we take into consideration that the autism dataset only had two categories, but the Alzheimer’s disease dataset had three categories. The classification results do not compare to the original study, but this can most likely be attributed to the lack of image preprocessing, because our limited analysis with the preprocessed dataset, which was class-imbalanced and had a very small sample size, showed better preliminary classification results. In conclusion, our findings emphasize the difficulty of Alzheimer's disease classification, and the importance of image preprocessing for medical imaging analysis. In the future, it would be interesting to obtain datasets of raw structural MRI images of the three original diseases we aimed to investigate, and preprocess them all in the same manner, including spatial normalization in order to conduct our analyses.

For our project, our ethical concerns are keeping the privacy of the patients intact. The datasets used were anonymized based on strict privacy standards such as the ABIDE dataset that confirms that their data “were fully anonymized (removing all 18 HIPAA protected health information identifiers, and face information from structural images). All data distributed were visually inspected prior to release.” (Di Martino, 2014).

# 7 – Contributions

Kangou implemented logistic regression models for both datasets.

Iyanu trained the ADD-Net model and applied KNN and SVM to both datasets.

Kyna contributed by exploring the autism phenotypic data, converting DICOM ADNI images to NIfTI format, processing the ADNI images through the model, and analyzing the classification models from the study and the preprocessed Alzheimer’s disease dataset.

Sandrine processed the ABIDE dataset through the model, did the exploratory analysis and dimensionality reduction techniques for both datasets, and conducted hypothesis testing and LDA for the Alzheimer dataset.

Each person wrote the section of the report corresponding to their own analysis.

# References

Ariadna Besga, Manuel Graña, & Darya Chyzhyk. (2020). Alzheimer's Disease versus Bipolar Disorder versus Health Control MRI data and processed results [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.3935636>.

Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Müller RA, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014 Jun;19(6):659-67. doi: 10.1038/mp.2013.78. Epub 2013 Jun 18. PMID: 23774715; PMCID: PMC4162310.

Fareed, M. M. S., Zikria, S., Ahmed, G., Mui-Zzud-Din, Mahmood, S., Aslam, M., Jillani, S. F., Moustafa, A., & Asad, M. (2022). ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans. *IEEE Access*, *10*, 96930–96951. <https://doi.org/10.1109/ACCESS.2022.3204395>

Korey P. Wylie and Benzi M. Kluger and Luis D. Medina and Samantha K. Holden and Eugene Kronberg and Jason R. Tregellas and Isabelle Buard (2023). Parkinson's disease, functional connectivity, and cognition. OpenNeuro. [Dataset] doi: doi:10.18112/openneuro.ds004392.v1.0.0

Liu, S., Masurkar, A. V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., Fernandez-Granda, C., & Razavian, N. (2022). Generalizable deep learning model for early Alzheimer’s disease detection from structural MRIs. *Scientific Reports*, *12*(1), 17106–17106. <https://doi.org/10.1038/s41598-022-20674-x>

Liu, S., Yadav, C., Fernandez-Granda, C. &amp; Razavian, N.. (2020). On the design of convolutional neural networks for automatic detection of Alzheimer’s disease. <i>Proceedings of the Machine Learning for Health NeurIPS Workshop</i>, in <i>Proceedings of Machine Learning Research</i> 116:184-201 Available from <https://proceedings.mlr.press/v116/liu20a.html>.

M Muslim, Ali (2022), “Brain MRI Dataset of Multiple Sclerosis with Consensus Manual Lesion Segmentation and Patient Meta Information”, Mendeley Data, V1, doi: 10.17632/8bctsm8jz7.1

Mohamad Abbass and Greydon Gilmore and Alaa Taha and Ryan Chevalier and Magdalena Jach and Terry M. Peters and Ali R. Khan and Jonathan C. Lau (2023). London Heath Sciences Center Parkinson's Disease Dataset (LHSCPD). OpenNeuro. [Dataset] doi: doi:10.18112/openneuro.ds004471.v1.0.1

Trevor K. M. Day and Tara M. Madyastha and Peter Boord and Mary K. Askren and Nicholas J. Wapstra and Thomas J. Montine and Thomas J. Grabowski (2023). ANT: Healthy aging and Parkinson's disease. OpenNeuro. [Dataset] doi: doi: 10.18112/openneuro.ds001907.v3.1.0