

# Alzheimer disease and Machine Learning

Théo AUDURIER

*La Rochelle Université, 17000 La Rochelle, France*

## ARTICLE INFO

Published on February 2024

**Keywords:**

Data analysis  
Alzheimer Dementia  
AD  
OASIS dataset  
Logistic classifier  
Machine Learning

## ABSTRACT

Between 1997 and 2050, the elderly population will increase from 63 to 137 million in the America (around 41% of the total 2021 population in America) and from 113 to 170 million in Europe (around 23% of the total 2021 population in Europe). About 1 in 9 people (10.8%) age 65 and older have Alzheimer's disease while people aged 75 to 84 are 13.1% to have Alzheimer's disease. Alzheimer disease is an incurable and progressive degenerative disease of the brain. Thus, the earlier a patient is diagnosed with Alzheimer's disease, the sooner we can support them. This study aims to find how can machine learning be useful to help medical staff with demented patients. To achieve this goal, we used the [Open Access Series of Imaging Studies \(OASIS\)](#) dataset. Thus, we analysed features from this longitudinal MRI data in nondemented and demented older adults and trained classification models to predict whether a patient has Alzheimer's disease. Finally, we observed that the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR) are both the most important features. According to the data, the MRI delay, the eTIV (Estimated intracranial volume) and the socio-economic status seems to play an important part in the model. However, to achieve a better accuracy of the patient's mental state, would an image classification algorithm be better e.g., an MRI image classifier?

## 1. Introduction

Alzheimer is sadly one of the most popular diseases for the elderly generations. This disease tends to push these unfortunate sufferers towards a state of dementia where they lose their episodic memory and their semantic memory (memory about general knowledge of world). But, unlike some others popular illness, not enough is known about Alzheimer. Indeed, this disease directly disrupts the brain. In the recent years, great advances have been made in the area of neuroimaging, genetics and diagnosis. However, the cause of Alzheimer's disease is still unclear. Some hypotheses tend to explain that an unusual accumulation of beta-amyloid protein could be one of the disease's triggers. Indeed, an alteration of this protecting protein could be in charge of the

disease appearance. These plaques lead to neurofibrillary degeneration, i.e. the death of neurons due to the interruption of intracellular transport. One other trigger could be a fault of the neurotransmitter production system: the acetylcholine. This neurotransmitter could play an important part in memory and learning. However, these hypotheses are questioned because they are not strongly correlated with the severity or the appearance of Alzheimer's disease. Alzheimer is also a question of rapidity, the earlier a patient is declared demented, the earlier this patient could be accompanied by medical staff. Some metrics has been implemented in order to monitor with potential Alzheimer's disease e.g., the Clinical Dementia Rating (CDR) which is a metrics to measure six cognitive domains, the Mini-

Mental State Examination (MMSE) which denotes the severity of cognitive impairment. However, technology tends to become more useful to help for medical discoveries. Nowadays, we can implement several algorithms to help research or medical advances. In this study, we will look if a machine learning algorithm could predict a potential Alzheimer's patient into one of these two groups: dement or non-dement. First of all, we will search for a consistent dataset with metrics like CDR or MMSE. Then, we will analyse these data to see if they fit well to this study or if we could find some relations between our metrics and the predictive feature. Afterwards, we will implement classification algorithms which will use the dataset to predict the patient's group. Next, we will look at the optimisation of these algorithms. Finally, we will see which information is the most important to predict a potential patient.

## 2. Methods

Data analysis requires a rigorous approach.

First and foremost, we need to dig deeper into the subject. Here, the goal is to discover how can machine learning be useful for Alzheimer's patients. However, machine learning algorithms require data. It is crucial to take a step back and to see if first, data are relevant secondly, if data come from trusted sources and then if the usage of these data are reasonable. Indeed, we could think of machine learning being a miracle algorithm that solve all problems but, this is not. Each dataset requires all above approach to be efficient. Thus, we defined the following guidelines: find out more about the hypothesis causes of Alzheimer's disease, define what we search for, find out a relevant dataset, analyse the features of the dataset, find out some features importances to predict the classification, choose some classification algorithms, compare the previous results of classification algorithms with fine-tuned hyperparameters of these previous algorithms, choose the best of these algorithms with his best parameters and finally, conclude.

In order to work properly, we used Jupyter notebook to leave a record of your work while annotating data and graph. Jupyter notebook uses Python programming language which is appreciated for his efficiency and all of his pre-built algorithms in Machine Learning and in Deep Learning. We used scikit-learn library for Machine Learning algorithms, seaborn and matplotlib

for visualization and Pandas to handle data. We also used Kaggle for the vast of dataset available for Machine Learning algorithms.

To measure the precision of all of the algorithms, we firstly separated the data in two variables: a train variable and a test variable. This split ensures the accuracy of the models. Indeed, the data could do a perfect score if the model was tested on his own trained data. Then, we created our own scoring function that print the confusion matrix and the accuracy score. The confusion matrix is really useful to see if the algorithm bad predicted a patient to a demented group or a non-demented. For human and safety reasons, we cannot allow a data model to classify an Alzheimer patient as not having dementia. The reverse situation would be easier to manage medically.

## 3. Results

### A. Dataset analysis

For this research project, we have selected a longitudinal collection (collection of repeatedly data from the same persons over an extended period of time) from the [Open Access Series of Imaging Studies \(OASIS\)](#) project.

We determined that the mean age is 77 years old with a max of 98. The first quartile equals to 71 years old and the third quartile equals to 82 years old. To describe further the dataset, we plotted the boxplot of the most relevant features to profile our dataset's patients.

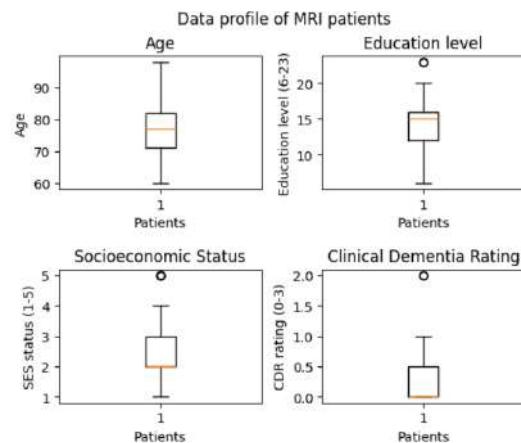


Figure 1. Data profile of MRI patients from the OASIS dataset.

Based on these data, we can notice that the socio-economic average is around 2 out of 5. We also can observe a education level

average around 15 out of 23. The dataset only contains patients from 0 CDR (clinical dementia rating) to 2 CDR. The CDR scale ranges from 0 to 3.

We also need to see the gender distribution in the dataset.

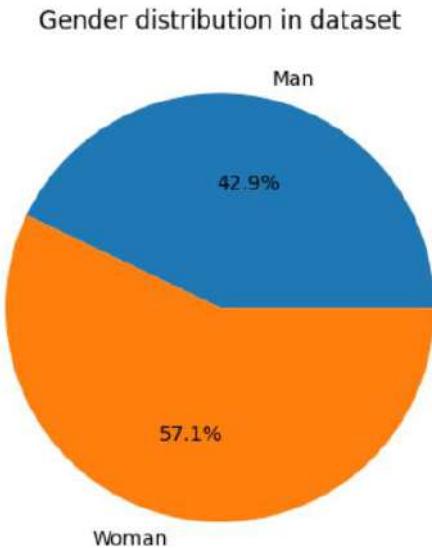


Figure 2. Gender distribution in the OASIS dataset.

This graph shows the gender distribution in the dataset. From the pie chart, the majority gender in the dataset seems to be woman. There is 14.2% less man in the dataset than woman. The number of men does not really seem to be in the minority compared to the number of women, even if the number of women remains high.

We now know the profile of our data. However, one problem could call into question the reliability of the classification algorithm. We need to see if gender is evenly or almost evenly distributed among demented

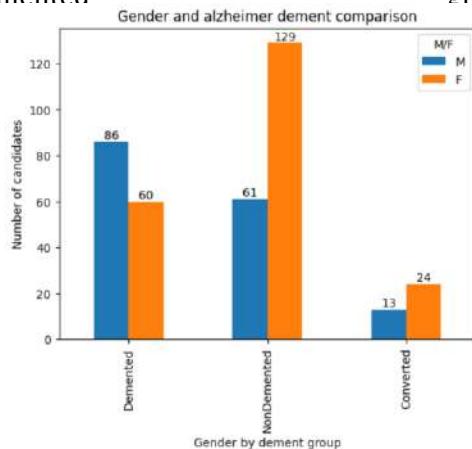


Figure 3. Gender and Alzheimer dement comparison

Based on this bar graph, the majority of demented patients seems to be men (around 59% of the demented group). In contrast, woman depict to be 68% of the non demented group distribution.

After cleaning the hand feature, which was only distributed by one value, we converted categorical values to binary ones. For the gender, we converted M to 0 and F to 1. For the group (predictive attribute), we converted 'Demented' to 1 and 'NonDemented' to 0. All 'converted' were pushed in demented's group.

Then, we looked at the missing values. We decided to dig further the relationship between SES and the predictive feature.

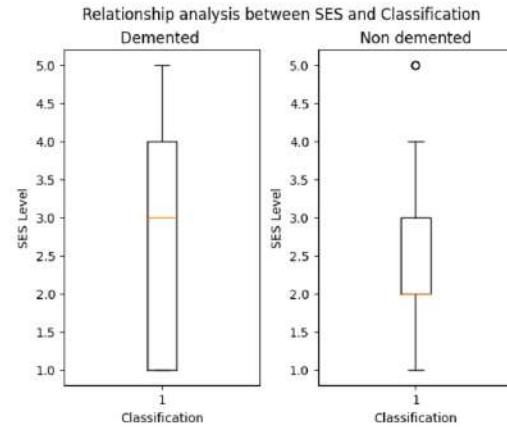


Figure 4. Relationship between SES and Classification.

The demented group have in average 3 out of 5 socio-economic level and a third quartile of 5 out of 5. The non demented group have a smaller distribution with a mean of 2 out of 5 and a third quartile of 4 out of 5.

We did the same for the MMSE (note to denotes the severity of cognitive impairment).

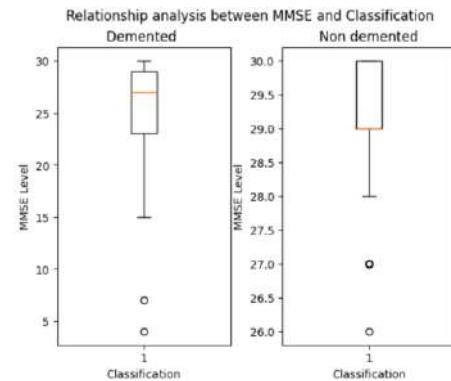


Figure 5. Relationship analysis between MMSE and Classification



Figure 6. Features heatmap

The mean of the MMSE level in the demented group is in average 27 out of 30. The first quartile is around 15 out of 30. The mean of the MMSE level in the non demented group is in average 29 out of 30. The first quartile is around 28 out of 30.

To determine the importance of features to predict the patient group, we need to plot a heatmap containing all features. ([Figure 6](#))

We can see that the two most important features to predict the group are the CDR score and the MMSE score. The higher the MMSE score, the less likely the patient is demented. In contrast, the higher the CDR score, the more likely the patient is demented. In the other side, the higher nWBV value is, the less likely the patient is demented.

### B. Machine Learning algorithms implementation

After analysing and cleaning the dataset, we implemented 6 classification algorithms: the Logistic Regression, the K-Nearest Neighbors (KNN), the Support Vector Machine (SVM), Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier. We got the following results without fine-tuning the model's hyperparameters.

Algorithm	Train Accuracy	Test Accuracy
Logistic Clf	.95	.902
KNN Clf	.69	.518
Support Vector Clf	.96	.920
Decision Tree Clf	1	.902
Random Forest Clf	1	.911
Gradient Boosting Clf	1	.902

*Figure 7. Classification algorithms scores without fine-tuning*

To compute these results, we used our own accuracy function ([See methods](#))

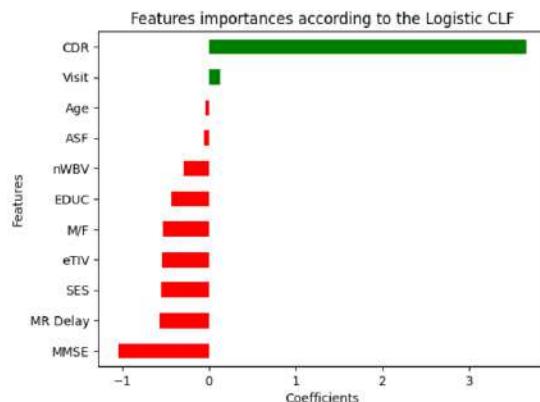
Then, we compared the following results of model's hyperparameters optimised.

Algorithm	Train Accuracy	Test Accuracy
Logistic Clf	.96	.919
KNN Clf	.93	.871
Support Vector Clf	.96	.919
Decision Tree Clf	.96	.919
Random Forest Clf	0.96	.919
Gradient Boosting Clf	1	.902

*Figure 8. Accuracy of models fine-tuned.*

We can see a lot of accuracy improvements. The KNN Classifier got better after standardisation and fine-tuning. So do the other algorithms.

Finally, we compared the features importances according to our logistic classifier algorithm.



*Figure 9. Features importances according to the logistic classifier.*

We can see that both CDR and MMSE help a lot to predict the predictive feature. In the other side, the MR Delay, the SES, the eTIV (standardised volume of the patient's cranial cavity) and the gender seems to be important according to the algorithm.

### C. Dataset interpretations

The age distribution in this data go from 60 years old to 98. Thus, the data only contains elderly people. The algorithm would probably miss sort young people. The solution to fix this problem would be to add more data from young people.

In the other hand, the mean of the socioeconomic status is around 2 out of 5. The third quartile is around 4 out of 5. The dataset is probably not enough

representative. This lack of uniformity could be a little issue for our classification algorithm.

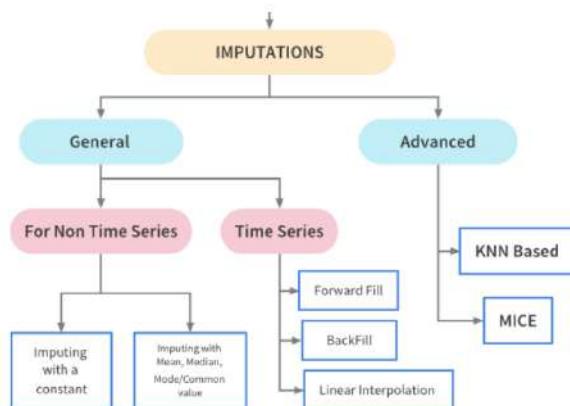
However, when we look at the MMSE ((note to denotes the severity of cognitive impairment) we can see that almost all non demented people have a MMSE score above 28 out of 30 or, almost all demented people have a MMSE between 30 and 15. This analysis approve of the MMSE scoring's duty. Indeed, the MMSE score the mental state of patients. A MMSE higher than 20 means a superficial mental trouble state. A MMSE between 10 and 20 means a moderate mental trouble state and below 10 highlight a severe mental trouble state. Thus, our analysis is in concordance with the MMSE use.

We also looked up at the differences between the gender and the predictive group. We saw that the number of women and man is not proportional inside our dataset. The main hypothesis behind that trouble is the lack of uniformity with our dataset. To improve the accuracy of our model, we should retrieve more men non demented data and more demented woman data.

On the heatmap ([Figure 6](#)) both MMSE and CDR seems to be important to predict the patient's dementia group. Both are metrics used by professional to evaluate the patient's mental state. Thus, our analysis is in concordance with these metrics.

#### D. Data cleaning process

In order to clean up data we used the K-Nearest Neighbor to fill in the empty values with the nearest coherent value. Indeed, we add two possibilities. Either we delete the empty data or we impute the empty data. Deleting the empty data would have lost 5% of our data. Considering that the loss is big, we imputed the empty data.



*Figure 10. Imputations data options for missing values*

Considering the [Parul's Pandey](#) diagram, we had the following coherent choices: either imputing with a median value or use an advanced option like KNN. If we had used the median value option, we would certainly have edited the importance of the feature with missing values. So, we choose the KNN algorithm.

#### E. Classification algorithm interpretations

With the library SKLearn we implemented 6 classifications algorithms: the Logistic classifier, the KNN Classifier, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier and the Gradient Classifier. On the first results table, almost all of our models performed really well (>90% d'accuracy on test data) (see [figure 6](#)). However, the K-Nearest Neighborhood did not perform well. His lack of efficiency is due to the fact that we have not standardized the values. After standardization and fine-tuning hyperparameters, the KNN improved itself by 35% (see [figure 7](#)).

The best algorithm to predict the patient's Alzheimer group would probably be the Gradient Boosting classifier despite his lower accuracy than the others models (except the KNN one). Indeed, his confusion matrix seems to be the best for what we need. We search an algorithm that combine the highest accuracy and the lowest wrong predicted negative values. The gradient boosting classifier has the lowest wrong predicted negative values (7) compared to (8) for the Random Forest, Decision tree, support vector machine or logistic regression.

## 4. Conclusions:

Despites that the dataset lacks uniform data like woman in demented group, man in non-demented group, we were able to identify some interesting facts. Based on our model's feature importances, higher the clinical dementia rating (CDR) is, the most probable the patient is demented. In the other hand, higher the Mini Mental State Examination (MMSE) is, the least probable the patient is demented. We also know that the MRI scan delay, the socio-economic status, the eTIV (standardised volume of the patient's cranial

---

cavity) can hardly affect the patient's probability to be demented. Machine Learning algorithms can be a great help in moving forward progress. We have also seen that standardisation and fine-tuning hyperparameters greatly improves model's accuracy. Ultimately however, we need a lot of well-distributed data to improve the prediction's accuracy and, as MRI scan becomes quicker thanks to technological advance should we better look at Alzheimer's predictions with patient's MRI images?

## Bibliography

- bank, T. w. (n.d.). *Total America population*. Retrieved from <https://data.worldbank.org/indicator/SP.POP.TOTL?end=2021&locations=US&start=2021>
- Pandey, P. (2020). *Parul Pandey's missing values diagram*. Retrieved from <https://www.kaggle.com/code/parulpandey/a-guide-to-handling-missing-values-in-python>
- Research, W. U. (n.d.). *Open Access Series of Imaging Studies 1, Longitudinal MRI Data in Nondemented and Demented Older Adults*. Dataset. Retrieved from <https://www.oasis-brains.org/>
- Worldometer. (2021). *Europe population*. Retrieved from <https://www.worldometers.info/world-population/europe-population/>