Building and analysing models for early prediction of Alzheimer's disease (AD)

Bojan Bogdanovic, 155002
Introduction to Bioinformatics
Faculty of Computer Science and Engineering
Skopje, 2020
Mentor: Ph.D. Monika Simjanoska

***Abstract*—In this project, Alzheimer's disease will be analyzed, which is a disease that belongs to the group of neurodegenerative diseases and is considered one of the most destructive and severe diseases of the human nervous system. The problem is how to diagnose it at the earliest possible stage before specific symptoms begin to appear. The main idea is to build an intelligent system that will be able to answer, based on certain biomarkers[1] from the subject, whether the disease is present or not.**

I. Introduction

Alzheimer's disease is the most common cause of dementia. It takes its name from the psychiatrist Alois Alzheimer, who in the early twentieth century was the first person to describe the disease. Over time, people who have Alzheimer's lose their memory and ability to concentrate. Orientation in space and time become increasingly difficult, and it is also harder for them to manage on their own in everyday life. Those affected need more support as the disease progresses.

The course of Alzheimer's can be positively influenced by a number of different medications and non-drug treatments – but it is not possible to cure the disease or to keep it from progressing. This makes it even more critical for people with Alzheimer's to receive good care and support.

There is no treatment that cures Alzheimer's disease or alters the disease process in the brain. In advanced stages of the disease, complications from severe loss of brain function — such as dehydration, malnutrition or infection — result in death. [1]

II. Symptoms

Most people mainly associate Alzheimer's disease with forgetfulness. But it can cause many different symptoms and develop in very different ways. How someone reacts to having Alzheimer's and copes with the disease not only depends on the changes in their brain, but also on that individual's personality, life experiences, current circumstances and relationships with other people.

A. Memory and cognitive ability

Most people's memory and other cognitive abilities gradually get worse as they get older. No longer being able to react as quickly and flexibly to new situations is a natural part of aging.This is different in people who have Alzheimer's disease. Their memory gradually fades away. At first, short-term memory is affected more. This means that they forget about events that have just happened, but can still remember experiences from long ago. But long-term memory also fades with time. The ability to concentrate is affected too, making it more and more difficult to maintain orientation in time and space.

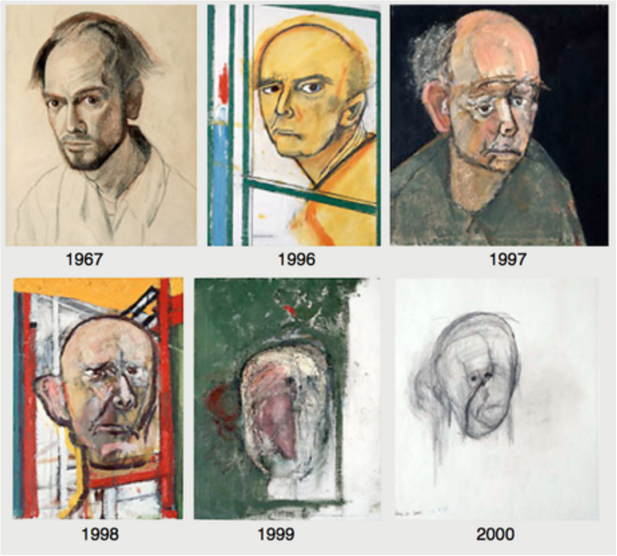


Fig. 1. Self-Portraits of an artist with Alzheimer's disease

B. Speech and language

We all have to search for the right word or feel tongue-tied now and then. But forgetting individual words more and more frequently is a different matter. As dementia progresses, it becomes more difficult to remember the right words, and people use words or phrases that do not match the context instead. This makes it difficult for others to understand them. And people with dementia also forget the meaning of words and are then often no longer able to follow conversations. This makes it increasingly harder to communicate verbally.

[1] measurable indicator of the severity or presence of some disease state. More generally a biomarker is anything that can be used as an indicator of a particular disease state or some other physiological state of an organism.

C. Mental health and changes in behavior

Many people with Alzheimer's go through noticeable changes in their behavior. Later on their personality may also change considerably. They can become unusually fearful, distrustful or passive, or may become aggressive as well. These changes can happen suddenly, and might cause fits of rage – or they may develop gradually. After all, a person who has Alzheimer's keeps finding themselves in situations that are confusing and in which they behave "wrongly." [2]

III. Causes

In Alzheimer's disease, more and more brain cells are lost as time goes by. It isn't clear why this happens. One thing that is known is that people with Alzheimer's don't have enough of an important chemical messenger called **acetylcholine** in their brain. And it has also been shown that small protein particles (*for example plaques*) build up in their brain. These might cause the nerve cells to die.

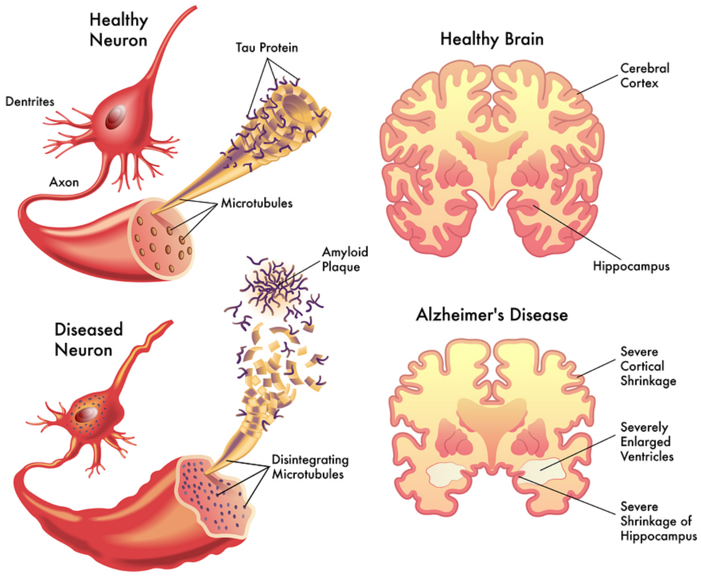


Fig. 2. Comparison between healthy and AD brain

But it is still not known what actually causes Alzheimer's disease. Several factors probably play a role.

IV. Risk factors

The risk of developing Alzheimer's increases with age, starting at about 65. Many studies have looked at whether particular life circumstances, diseases or behaviors can increase or lower the risk of Alzheimer's. There are several important risk factors that are known to cause dementia. The **alipoprotein E4** variant (APOE E4) is a gene that is the largest known risk factor for AD. Subjects with APOE E4 have a risk 10 to 30 times higher of developing AD compared to non-carriers (*i.e. subjects without the gene*). The exact mechanism through which the presence of APOE E4 leads to AD is not known.

Another known and important risk factor for AD is **age** – the older subjects are the more likely they are to develop AD. Above the age of 65, the risk of developing dementia doubles every 5 years. Gender is another known risk factor, where women seem more likely to develop AD than men. The reasons for this are still unclear. [3]

Finally, there exist many other risk factors related to existing medical conditions and lifestyle. Medical conditions such as **type 2 diabetes**, **high blood pressure**, **high cholesterol**, **obesity** or **depression** are known to increase the risk of developing dementia. Lifestyle factors known to increase the risk of developing dementia include **physical inactivity**, **smoking**, **unhealthy diet**, **excessive alcohol** or **head injuries**.

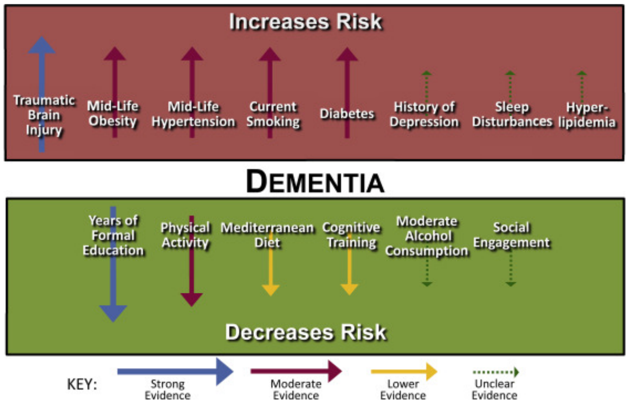


Fig. 3. Diagram showing different risk factors related to lifestyle and the associated level of evidence

V. Diagnosis

It is not yet possible to diagnose Alzheimer's with complete certainty using the currently available tests while the person is still alive. The disease is diagnosed if someone has the typical symptoms and no other cause can be found. Looking at the brain using imaging techniques like **computed tomography** (CT) or **magnetic resonance imaging** (MRI) is not enough to tell whether or not someone has Alzheimer's disease.

Because symptoms like forgetfulness, changes in behavior and problems with orientation can have many different causes, it is important not to rush to a diagnosis of Alzheimer's. The symptoms might also be caused by depression or other physical conditions like meningitis, a stroke or bleeding in the brain.

VI. Defining the problem

Techniques to predict onset of AD are badly needed to help focus trials of potential treatments on the right people. To conduct an effective clinical trial it is crucial to:

1) Identify individuals that are likely to need and respond to treatment. AD treatments are most likely to be effective at early disease stages, even before any outward signs of dementia.
2) Accurately predict the change in disease indicators so that we can assess the effect of the treatment.

Our limited understanding of AD makes prediction of symptom onset hard. However, several approaches are available in the scientific literature:

- **Manual prediction by a clinical expert**
  An informed clinician experienced in interpreting multi-modal data can judge prognosis and predict conversion to a more severe diagnostic category by drawing on their knowledge of the clinical history of patients with a similar presentation, e.g. through visual rating of brain scans.

Fig. 4. Clinical dementia expert Professor Nick Fox explains to the former prime minister of the UK, David Cameron, how to spot signs of Alzheimer's disease in brain images.

- **Statistical prediction using regression**
  Regression is a statistical technique to model the relationship between variables and thus to predict one set of variables from another. One might regress markers of AD or clinical assessments against time in historical data to make predictions of future measurements or changes in patient status. Examples from the literature include regression of clinical diagnosis against anatomical volumes from MRI [4], cognitive test scores [5], rate of cognitive decline [6] and retrospectively staging subjects by time to conversion between diagnoses [7].

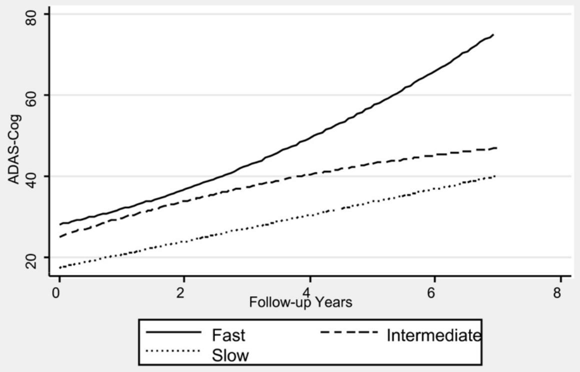


Fig. 5. Predicted ADAS-Cog values for different patient subgroups (slow, intermediate or fast progressors, determined by their rate of decline in MMSE score) estimated using statistical regression.

- **Machine learning**
  Supervised machine learning techniques, such as support vector machines, random forests, and artificial neural networks, learn the relationship between the values of a set of predictors and their labels. They can prove very effective in high dimensional classification and regression problems. In AD, Klöppel et al. [8] showed the ability to discriminate AD patients from cognitively normal subjects from MR images using support vector machines. Later work by Zhang et al. [9] uses a wider variety of biomarkers.

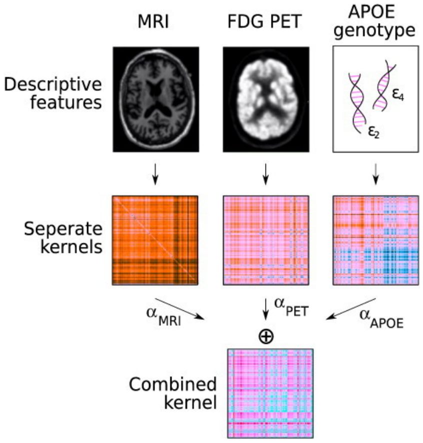


Fig. 6. An example of a supervised machine learning technique combining multi-modal patient information to make a prediction.

- **Data-driven disease progression models**
  Data-driven disease progression models are a more recent innovation in AD modelling and prediction using unsupervised learning. They do not rely techniques on prior knowledge of disease status, but rather aim to extract a picture of how all biomarkers evolve concurrently during the disease. Examples include models built on a set of scalar biomarkers to produce discrete [10] [11] or continuous [12] [13] pictures of disease progression; richer but less comprehensive models that leverage structure in data such as MR images [14] [15].

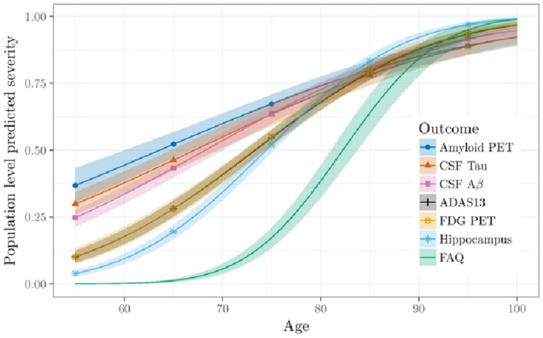


Fig. 7. Predicted severity of different biomarkers with age estimated using a data-driven disease progression model.

Early forecasting of the AD diagnosis is very crucial for patients to start being treated at earliest stage. This can help in slowing down progression of the disease and symptoms as well. In this project, the goal is to build several different models that will classify subjects based on their diagnosis into three categories - cognitively normal (CN), mild cognitive impairment (MCI) or Alzheimer's disease (AD). Models will be trained and tested on same dataset and afterwards used to predict diagnosis of unknown subjects. They will be qualitatively compared (*e.g. their accuracy*), concluding which one did a better job. From the huge set of features, only small most relevant subset will be extracted in order to simplify the solution of the problem.

VII. Description of the data set

The data set that will be used in this project is taken from **ADNI** (*Alzheimer's Disease Neuroimaging Initiative*) under the name TADPOLE_D1_D2.csv. In order to be able to download a certain data set from them, one must send a request with an explanation for which purposes it will be used first. More information can be found on their website.

The original data set contains data from 12741 subjects divided by 1907 attributes. Mainly the attributes are divided into two categories: **quantitative biomarkers** (*e.g. MRI scans of different parts of the brain*) and **personal information** (*e.g. participant ID, demographic information, etc.*). Many of the attributes are **redundant**, meaning they do not contain data for most patients.

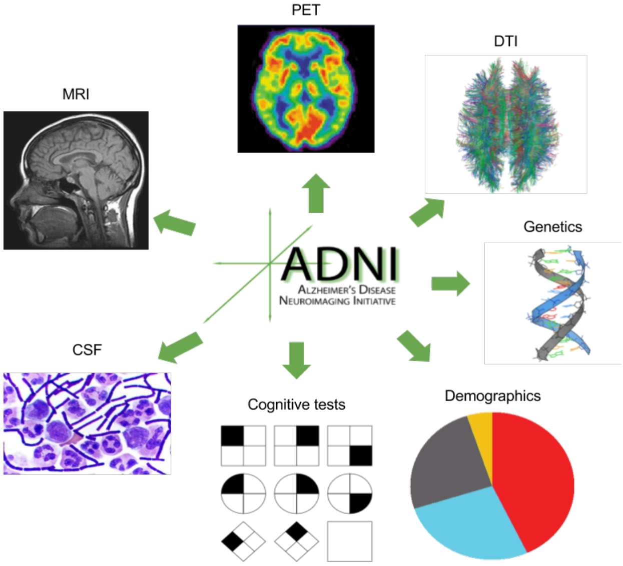


Fig. 8. Diagram showing original TADPOLE biomarkers

To simplify the problem, we shorten the data set by considering only a small subset of biomarkers that are known to be informative. Our reduced data set initially contains 17 columns and same number of rows as the original data set.

Each row represents data for one subject, and each column represents a feature or measurement (commonly called biomarker) from the subject. Our features can be divided into 6 categories:

- **Personal information**:
  - **PTID**: Participant ID
  - **AGE**: Age at baseline
  - **PTGENDER**: Sex
  - **PTEDUCAT**: Years of education
  - **PTRACCAT**: Race
- **Gene expression**:
  - **APOE4**: Expression of the ApoE4 gene
- **PET measures**:
  - **FDG**: measure cell metabolism, where cells affected by AD show reduced metabolism
  - **AV45**: measures amyloid-beta load in the brain, where amyloid-beta is a protein that mis-folds (*i.e. its 3D structure is not properly constructed*), which then leads to AD
- **MRI measures**:
  - **Hippocampus**: scan of a complex brain structure embedded deep into temporal lobe
  - **WholeBrain**: scan of the subject's whole brain
  - **Entorhinal**: scan of an area of the brain that is located in the medial temporal lobe and functions as a hub in a widespread network for memory, navigation and the perception of time
  - **MidTemp**: scan of the middle temporal artery
- **Cognitive tests**:
  - **CDRSB**: Clinical Dementia Rating Scale - Sum of Boxes
  - **ADAS11**: Alzheimer's Disease Assessment Scale 11
  - **MMSE**: Mini-Mental State Examination
  - **RAVLT_immediate**: Rey Auditory Verbal Learning Test (sum of scores from 5 first trials)
- **Target**:
  - **DX_bl**: Subject's diagnosis

DX_bl represents the **target variable** of which we want to gain a deeper understanding. We will try to build models whose goal is to predict the value of this variable based on the values of other features. In other words, we will try to forecast subject's diagnosis analysing subject's biomarkers mentioned above.

The target variable can result in any of the following five values: **CN** (*Cognitive Normal*), **EMCI** (*Early Mild Cognitive Impairment*), **LMCI** (*Late Mild Cognitive Impairment*), **SMC** (*Significant Memory Concern*) and **AD** (*Alzheimer's Disease*)

VIII. Description of the data

For advancing the diagnosis of dementia, assessment of quantitative biomarkers (medical measurements that can indicate a disease) in addition to cognitive tests is of great value. In order to gain a better understanding of our data, their nature and significance, a more detailed explanation of each category of biomarkers follows.

A. Cognitive Tests

Cognitive tests are **neuropsychological tests** administered by a clinical expert which assess several skills: general cognition, memory, language, vision, etc.. These cognitive tests give an overall sense of whether a person is aware of their symptoms, is aware of the surrounding environment (*i.e. he/she knows where they are, know the date and time*) and whether he/she can remember a short list of words, follow instructions and do simple calculations.

Cognitive tests are important in Alzheimer's disease because they measure cognitive decline in a direct and quantifiable manner. In the cascade of pathological events that lead to Alzheimer's disease, cognitive decline is one of the latest to become abnormal. This is because the first abnormalities are first noticed on the microscopical scale through the misfolding of a protein called Amyloid beta. These are followed by changes at larger scales: loss of the neurons myelin sheath, neuron death, visible atrophy in MRI scans and finally cognitive decline.

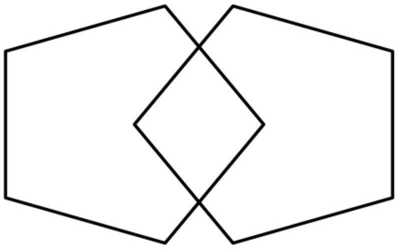


Fig. 9. Cognitive tests can help in the diagnosis of AD. In the tests, people are instructed to copy drawings similar to the one shown in the picture, remember words, read and subtract numbers. These intercalated pentagons are used in the Mini-Mental State Examination (MMSE), an extensively used cognitive test.

These tests have several limitations:

1) They suffer from practice effects, i.e. patients who undertake the same test several times can learn/remember how to do it, and thus score higher at a follow-up visit; this limits the usefulness of the test in assessing dementia.
2) They have floor or ceiling effects, which means that many subjects might score the highest/lowest score possible.
3) They can be biased, as they are undertaken by a human expert who might be influenced by prior knowledge of the subject's cognitive abilities.

B. MRI Measures

Magnetic resonance imaging (MRI) is a technique used to image the anatomy and the physiological processes of the brain and other body parts. With MRI, atrophy can be quantified by measuring the volume of **gray matter** (GM) and **white matter** (WM) of the brain. The GM is the brain tissue that consists of nerve cells and the WM consists of fibres connecting these nerve cells. GM can be found in the cortex of the brain and in sub-cortical areas. As a structural MRI scan shows contrast (*i.e. differences in pixel intensities*) between these tissues, it can be used for volume measurement.

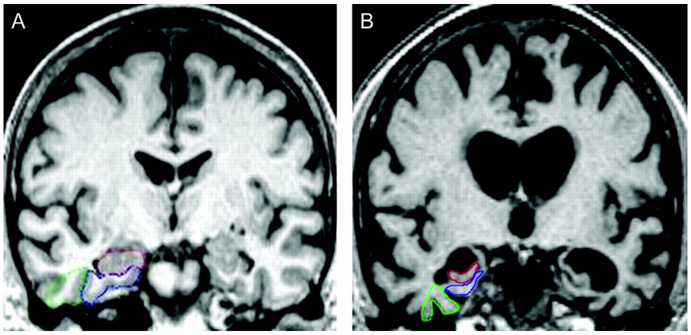


Fig. 10. Left: MRI scan of a subject before the onset of atrophy. Right: MRI scan of the subject with severe atrophy due to AD, which is visible throughout the brain. The coloured regions represent deep gray matter structures, affected early in the disease process (*hippocampus = red; entorhinal cortex = blue; perirhinal cortex = green*). MRI is a widely used technology for measuring the extent of atrophy and tracking the progression of Alzheimer's disease (AD).

Atrophy is indicated by the loss of volume in a particular brain region between two scans, one initial scan and one follow-up scan. Atrophy is caused by the death of neurons in regions affected.Quantification of atrophy with MRI is a very important biomarker as it is widely available and non-invasive. Also, it is a good indicator of progression of MCI to dementia in an individual subject because it becomes abnormal in close temporal proximity to the onset of the cognitive impairment

C. PET Measures

Positron Emission Tomography (PET) detects pairs of gamma rays emitted by a radioactive tracer, which is introduced into the body of a biologically active molecule. Three-dimensional images of tracer concentration within the body are then constructed by computer analysis. Before a PET scan, the patient is injected with a contrast agent (containing the tracer) which spreads throughout the brain and binds to abnormal proteins (amyloid and tau). This enables researchers to track the concentration of these proteins. PET scans can be of several types, depending on the cellular and molecular processes that are being measured:

- **Cell metabolism** using **Fluorodeoxyglucose (FDG) PET**: Neuronal cell metabolism refers to the the activity going on inside neuronal cells such as the processing of food and elimination of waste. Neurons that are about to die will show reduced metabolism, so FDG PET is an indicator of neurodegeneration. FDG PET can be used to measure cell metabolism.

- **Levels of abnormal proteins** such as **amyloid-beta through AV45 PET**: Amyloid-beta misfolding (i.e. errors in the construction of its 3D structure) is thought to be one of the causes of Alzheimer's disease. High levels of misfolded amyloid-beta in the brain are thought to eventually lead to future neurodegeneration and cognitive decline. AV45 PET can be used to measure the levels of amyloid in the brain.

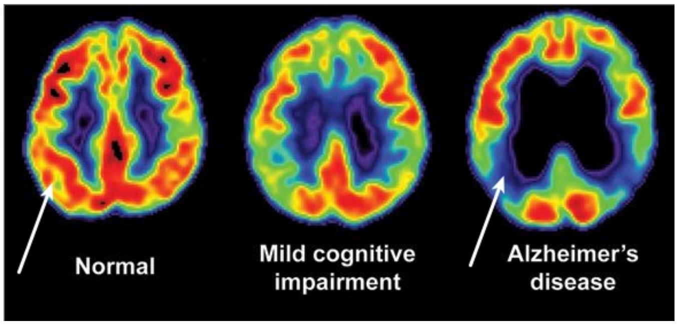


Fig. 11. Fluorodeoxyglucose (FDG) PET images for a cognitively normal subject (*left*), a subject with mild cognitive impairment (*middle*) and Alzheimer's disease (*right*). FDG PET measures cellular metabolism, which is known to decrease during the development of AD. There is decreased metabolism in the parietal region (*white arrow*) in the Alzheimer's subject compared to the cognitively normal subject.

The PET measures are important because they give information about molecullar processes that happen in the brain. These are usually the first to become abnormal in the cascade of events that lead to Alzheimer's disease, and are therefore important early markers of the disease that is about to unfold.

While PET scans are non-invasive, they have some limitations. One main limitations is that the patient is exposed to ionizing radiation, which limits the number of scans they can take in a specific time interval. PET scans also have a much lower spatial resolution compared to MRI scans.

D. Gene Expression

Gene expression measurement is usually achieved by quantifying levels of the gene product, which is often a protein. Two common techniques used for protein quantification include Western blotting and enzyme-linked immunosorbent assay or ELISA. However, the gene expression level can also be inferred by measuring the level of mRNA, which is achieved using a technique called Northern blotting.

The **APOE gene** provides instructions for making a protein called apolipoprotein E. This protein combines with fats (*lipids*) in the body to form molecules called lipoproteins. There are at least three slightly different versions (*alleles*) of the APOE gene. The major alleles are called e2, e3, and e4. The **e4 version** of the APOE gene increases an individual's risk for developing late-onset Alzheimer disease. People who inherit one copy of the APOE e4 allele have an increased chance of developing the disease; those who inherit two copies of the allele are at even greater risk. In our data set, we keep information about subject's number of e4 alleles (0, 1 or 2).

It is important to note that people with the APOE e4 allele inherit an increased risk of developing Alzheimer disease, not the disease itself. Not all people with Alzheimer disease have the APOE e4 allele, and not all people who have this allele will develop the disease.

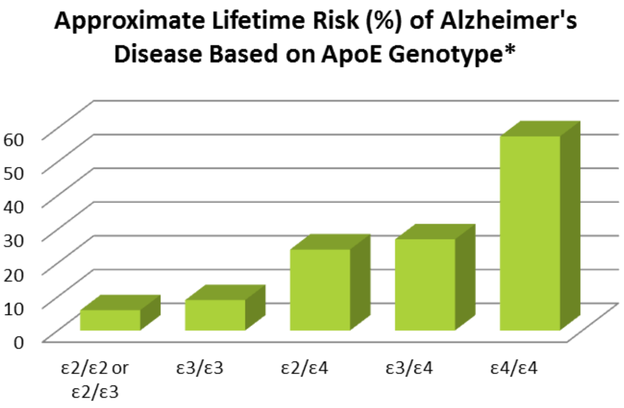
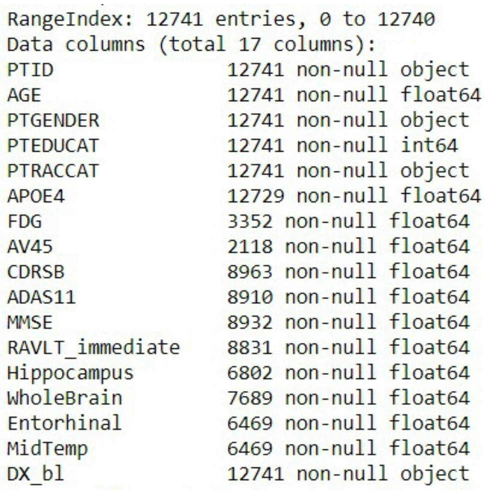


Fig. 12. These figures relate to ApoE-based genetic risk independently. There are other risk factors (*genetic or non-genetic*) that may modify the risk in an individual.

IX. Data Preprocessing

A. Incomplete data

First step in data preprocessing is to detect all subjects with incomplete data and to remove them from our data set. Incomplete data can cause a lot of troubles in the process of building and training our models later.



```
RangeIndex: 12741 entries, 0 to 12740
Data columns (total 17 columns):
PTID                  12741 non-null object
AGE                   12741 non-null float64
PTGENDER              12741 non-null object
PTEDUCAT              12741 non-null int64
PTRACCAT              12741 non-null object
APOE4                 12729 non-null float64
FDG                   3352 non-null float64
AV45                  2118 non-null float64
CDRSB                 8963 non-null float64
ADAS11                8910 non-null float64
MMSE                  8932 non-null float64
RAVLT_immediate       8831 non-null float64
Hippocampus           6802 non-null float64
WholeBrain            7689 non-null float64
Entorhinal            6469 non-null float64
MidTemp               6469 non-null float64
DX_bl                 12741 non-null object
```

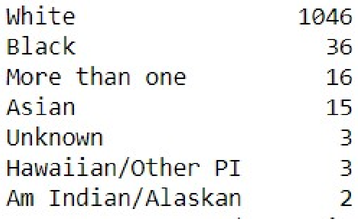
Fig. 13. Concise summary of the data within our dataset.

As it can be seen from the figure above, all subjects contain personal information data and diagnosis, but not all of them have data for the biomarkers.

For example, only 2118 subjects have data for the AV45 attribute. Removing all subjects (rows) that contain any missing data from our dataset results into new dataset with 1121 entries. In fact, our dataset contained 91% incomplete data.

B. Redundant features

In the next step, we want to remove any irrelevant feature, so we can simplify the dataset even more. If we take a closer look of values in the PTRACCAT column, we can see that almost 93% entries have value ’white’. It means that this attribute doesn’t give us enough information about possible racial predisposition for the disease. Most of the subjects belong to same racial group and only few belong to other groups, so this feature will be excluded.



```
White                1046
Black                  36
More than one          16
Asian                  15
Unknown                 3
Hawaiian/Other PI       3
Am Indian/Alaskan       2
```

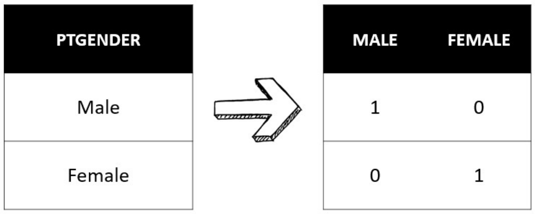
Fig. 14. Counts of unique values in PTRACCAT column.

We can also omit the PTID attribute because it is an identification number for each subject and has no meaning for the models we want to build.

C. Encoding categorical data

In Fig. 6, it can be seen that attributes **PTGENDER** (Male/Female) and **DX_bl** (CN/EMCI/LMCI/SMC/AD) are of type *object*. To be more precise, they contain strings and they both represent **categorical data**. Categorical data is the data that generally takes a limited number of possible values. Also, the data in the category need not be numerical, it can be textual in nature. All machine learning models are some kind of mathematical model that need numbers to work with. This is one of the primary reasons we need to pre-process the categorical data before we can feed it to machine learning models. There are different types of encoding schemes, the most common are: assigning 0/1 to binary data and assigning increasing integers to a category of related data (*e.g. temperature: low = 1, medium = 2, high = 3*).

Let’s consider **PTGENDER** attribute where ’Male’ is assigned the value as 1 while ’Female’ is 0. In all the calculations that are going to take place, the weight of Male is going to be more than that of Female. This does not make sense because Gender is a category of data and both variables need to be treated equally by the model to predict accurate results. That is why, for encoding the PTGENDER attribute, we are going to use **One-Hot encoding**. This encoding is appropriate for categorical data where no relationship exists between categories. It involves representing each categorical variable with a binary vector that has one element for each unique label and marking the class label with a 1 and all other elements 0. In other words, we will divide the PTGENDER column into two (MALE and FEMALE), assigning 1 to one column and 0 to the other, depending on subject’s gender.

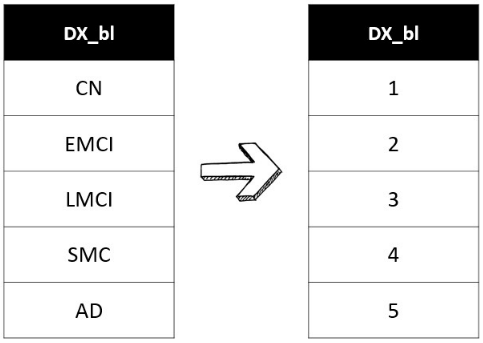


PTGENDER
Male
Female

MALE	FEMALE
1	0
0	1

Fig. 15. One-Hot encoding of PTGENDER attribute.

On the other hand, we can see that there is a certain order related to the values of the target attribute **DX_bl**. We can rank values, from CN to AD, based on the subject’s neuropsychological disorder. Here, we are going to use **Label encoding** where we simply convert labels to integer values in ascending order.



DX_bl
CN
EMCI
LMCI
SMC
AD

DX_bl
1
2
3
4
5

Fig. 16. Label encoding of DX_bl attribute.

D. Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. When there is correlation between two features, it means that one of them doesn’t provide us new information for our model, thus for better performances it can be omitted. Here we will analyse linear correlation between our features. **Linear correlation** is a measure of dependence between two random variables that can take values between -1 (*negative correlation*) and 1 (*positive correlation*), 0 means no correlation. Best way to analyse correlations between all features is using **heat map representation**. It must be mentioned that this type of correlation gives correct results only for continuous data. Another thing we must remember is that correlation does not imply causation.

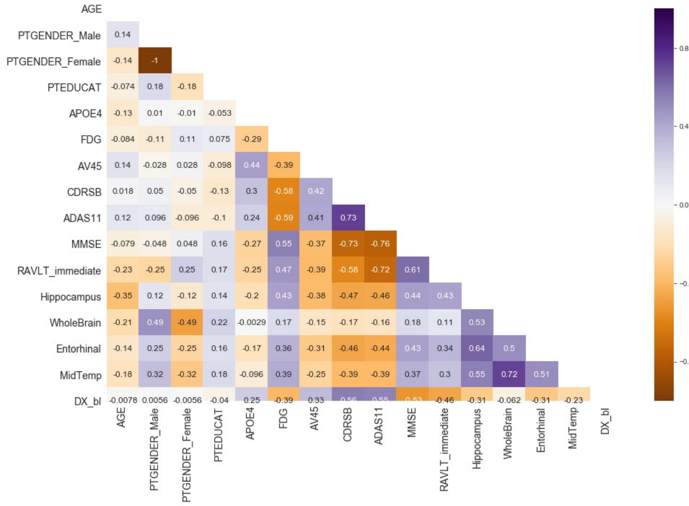


Fig. 17. Linear correlation heat map for the dataset.

First thing we can notice is -1 correlation between PTGENDER_Male and PTGENDER_Female.This coefficient doesn't tell us anything relevant because both features are discrete, so we will ignore it.

On the other hand, we can notice a trend of pretty high coefficients between ADAS11 and other cognitive tests results. In fact, highest negative coefficient is between ADAS11 and MMSE (-0.76) and highest positive coefficient is between ADAS11 and CDRSB (0.73). It is possible that this feature doesn't provide us any new information. It seems like it contains repetitive information from other tests. To determine the correlation, the coefficient alone is not sufficient. It is best to display the graph between the two features to better understand the dependency.

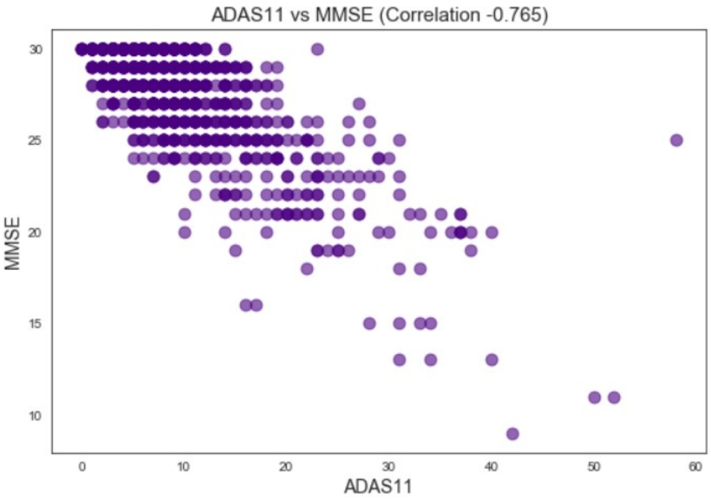


Fig. 18. Scatter plot between ADAS11 & MMSE

From the plot we can see that there really is some kind of linear dependency between these two features. Plots of ADAS11 and CDRSB/RAVLT_immediate show something similar too. Considering all of the above, we're assuming that ADAS11 doesn't provide us new information thus it is redundant for our model. It will be omitted from the final dataset.

This completes the data pre-processing phase. The dataset is ready to serve as a source for training and testing the models we want to build.

X. Choosing machine learning algorithms

Since our data is already labeled with the correct answer, our problem belongs to the group of **supervised learning** problems. Our target attribute is categorical, so we will be building **classification models**. We are going to teach or train our models using data which is already tagged with the correct answer. After that, models are provided with the new set of examples (*data*) without the target label (*desired output*) and models predict the output of the new data by applying their learning from historical trained data.

Choosing the most optimal algorithm for solving one particular problem depends on many factors like: size of the training data, accuracy and/or interpretability of the output, speed or training time, linearity, number of features, etc.. Our medium sized dataset with average number of features allows us to try experimenting with more complex algorithms. Here, two types of classifiers will be built using different machine learning algorithms: **Random Forest** and **Neural Network**.

A validation dataset is a dataset of examples used to tune the hyperparameters (*i.e. the architecture*) of a classifier. In order to validate our classifiers properly, we are using **cross-validation** - we are doing a sequence of fits where each subset of the data is used both as a training set and as a validation set. Here we split the data into five groups, and use each of them in turn to evaluate the model fit on the other 4/5 of the data. This is called **five-fold cross-validation**. As an accuracy measure, we take the mean from all trials.

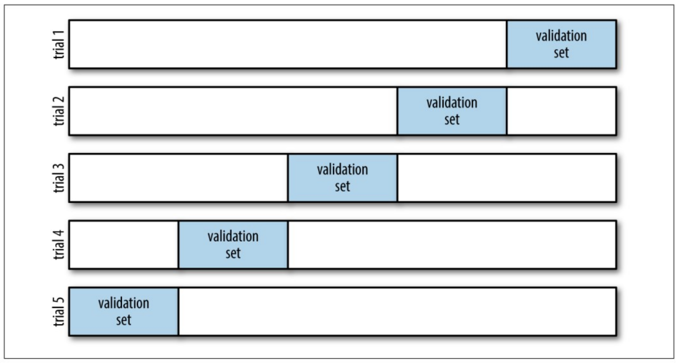


Fig. 19. Visualization of five-fold cross-validation.

One crucial step in building ML model is tuning it's **hyperparameters** - the arguments that can be set before training and which define how the training is done. These parameters are tunable and can directly affect how well a model trains. Thus, in order to achieve maximal performance, it is important to understand how to optimize them.

XI. Random Forest Classifier

Random Forest (RF) is one of the many machine learning algorithms used for supervised learning and it can be used for both classification and regression tasks. RF is based on **decision trees**. In machine learning decision trees are a technique for creating predictive models. They are called decision trees because the prediction follows several branches of "*if. . . then. . .*" decision splits - similar to the branches of a tree. RF makes predictions by combining the results from many individual decision trees - so we cal them a **forest** of decision trees.

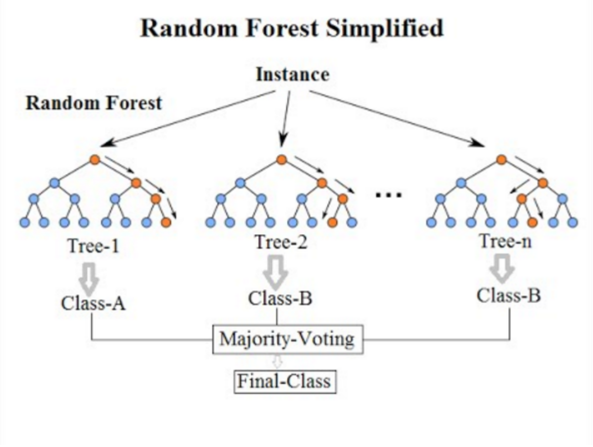


Fig. 20. Simplified version of the Random Forest diagram.

Single decision trees are very easy to visualize and understand because they follow a method of decision-making that is very similar to how we humans make decisions: with a chain of simple rules. However, they are not very **robust**, i.e. they don't generalize well to unseen samples. Here is where Random Forests come into play.

A. Hyperparameters

For finding the best combination of values for hyperparameters, we define a range of values for every parameter and then use **Grid Search** which evaluates all combinations we define and chooses the best one. The main hyperparameters in Random Forests are:

- The **number of decision trees** to be combined - in our model this number is **200**.
- The **maximum depth** of the trees - in our model the maximum depth is defined as **None**.
- The **maximum number of features** considered at each split - in our model this is defined as **auto**.
- Whether **bagging/bootstrapping**[2] is performed - for our model this is set to **True**, otherwise the whole dataset would be used to build each tree.
- The **minimum number** of samples required to **split an internal node** - in our model this parameter has a value of **6**.

XII. Neural Network Classifier

A **Neural Network** (NN) consists of **units** (*neurons*), arranged in **layers**, which convert an input vector into some output. Each unit takes an input, applies a (*often nonlinear*) function to it and then passes the output on to the next layer. Generally the networks are defined to be **feed-forward**: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. **Weightings** are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand. This is the learning phase.

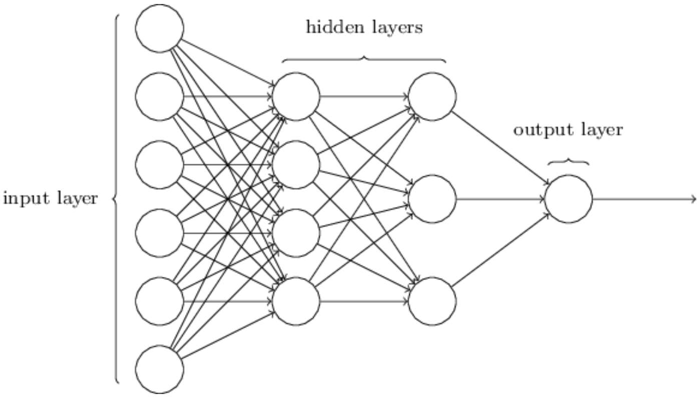


Fig. 21. Simplified version of the Neural Network diagram.

It is very effective for high dimensional problems, able to deal with complex relations between variables, non-exhaustive category sets and complex functions relating input to output variables. NN has powerful tuning options to prevent over- and under-fitting.

On the other hand, it is theoretically complex and difficult to implement (*although deep learning frameworks are readily available that do the work*). Neural networks are non-intuitive and require expertise to tune. In some cases requires a large training set to be effective.

A. Hyperparameters

Hyperparameter optimization is a big part of deep learning. The reason is that neural networks are notoriously difficult to configure and there are a lot of parameters that need to be set. On top of that, individual models can be very slow to train. Here, we will try tuning some of those parameters with the already known method - **Grid Search**. For the other parameters, we are going to take their default or most used values.

- **Number of Hidden Layers and units**: Hidden layers are the layers between input layer and output layer. Many hidden units within a layer with regularization techniques can increase accuracy. Smaller number of units may cause **under-fitting**. In our model, we are having one hidden layer with **12** units .

[2] Default method where decision trees are trained on randomly sampled subsets of the data, while sampling is being done with replacement.

- **Dropout Regularization**: Dropout is a regularization technique to avoid over-fitting (*increase the validation accuracy*) thus increasing the generalizing power. A probability too low has minimal effect and a value too high results in under-learning by the network. In our model, we are going to use dropout rate of **0.1**
- **Network Weight Initialization**: Ideally, it may be better to use different weight initialization schemes according to the activation function used on each layer. Mostly **uniform distribution** is used. We are going to use the same distribution in our model.
- **Neuron Activation Function**: Activation functions are used to introduce nonlinearity to models, which allows deep learning models to learn nonlinear prediction boundaries. Generally, the **rectifier activation function** is the most popular. **Sigmoid** is used in the output layer while making **binary predictions**. **Softmax** is used in the output layer while making **multi-class predictions**. Because our goal is to classify instances in more than two classes, for our output layer we use **Softmax**, and for our hidden layer we use **ReLU** function, which is most popular hidden-layer activation function.
- **Learning Rate**: The learning rate defines how quickly a network updates its parameters. **Low learning rate** slows down the learning process but converges smoothly. **Larger learning rate** speeds up the learning but may not converge. In our model, we are using optimizer called ’**Adam**’ which automatically tunes learning rate and momentum with every epoch.
- **Momentum**: Momentum helps to know the direction of the next step with the knowledge of the previous steps. It helps to prevent oscillations. A typical choice of momentum is between 0.5 to 0.9.
- **Number of Epochs**: This is the number of times the whole training data is shown to the network while training. The number of epochs should be increased until the validation accuracy starts decreasing, even when training accuracy is increasing (*over-fitting*). For our model, this parameter will have value of **100**.
- **Batch Size**: Mini batch size is the number of sub samples given to the network after which parameter update happens. In our model, we are going to use batch size of **128**.

XIII. Splitting the dataset

One of the first decisions to make before evaluating our models is how to utilize the existing data. One common technique is to split the data into two groups typically referred to as the **training** and **testing sets**. The **training set** is used to develop models and feature sets - they are the substrate for estimating parameters, comparing models, and all of the other activities required to reach a final model. The **test set** is used only at the conclusion of these activities for estimating a final, unbiased assessment of the model's performance.

In order to build and evaluate our classifiers properly, we are splitting our data set into most commonly used way. In fact, 70% of the data will be contained in the training set, while the remaining 30% in the testing set. Before splitting, the data is shuffled.

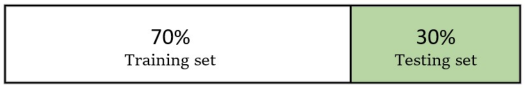


Fig. 22. Visualization of the train-test split.

XIV. Classification metrics

After doing the usual feature Engineering, selection, and of course, implementing a model and getting some output in forms of a probability or a class, the next step is to find out how effective is the model based on some metric using test datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms. We will be focusing on the ones used for classification problems.

A. Confusion Matrix

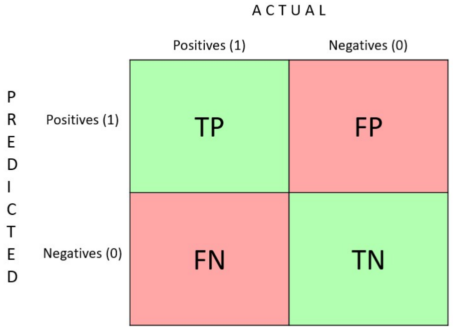
The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for classification problem where the output can be of two or more types of classes. We can consider the confusion matrix as a table with two dimensions and sets of ”classes” in both dimensions. Our actual classifications are columns and predicted ones are rows.

Because in our problem, the main goal is to determine whether a subject has an AD or not, we can simplify the representation to a binary level using ”one vs. others” technique.

Let's give a label of to our target variable:
1: When a person is having AD.
0: When a person is NOT having AD.
Now we can represent the matrix as a 2x2 table where:

- **True Positives (TP):** True positives are the cases when the actual class of the data point was 1 (*True*) and the predicted is also 1 (*True*). In our example, the case where a person is actually having AD (1) and the model classifying his case as AD (1) comes under True Positives.
- **True Negatives (TN):** True negatives are the cases when the actual class of the data point was 0 (*False*) and the predicted is also 0 (*False*). The case where a person NOT having AD and the model classifying his case as Not AD comes under True Negatives.
- **False Positives (FP):** False positives are the cases when the actual class of the data point was 0 (*False*) and the predicted is 1 (*True*). False is because the model has predicted incorrectly and positive because the class predicted was a positive one (1). A person NOT having AD and the model classifying his case as AD comes under False Positives.

- **False Negatives (FN):** False negatives are the cases when the actual class of the data point was 1 (*True*) and the predicted is 0 (*False*). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0). A person having AD and the model classifying his case as No-AD comes under False Negatives.



		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

Fig. 23. Simplified version of the Confusion matrix

The ideal scenario that we all want is that the model should give 0 False Positives and 0 False Negatives. But that's not the case in real life as any model will **not** be 100% accurate most of the times. The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion matrix and the numbers inside it.

B. Accuracy

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.In the numerator, are our correct predictions and in the denominator, are the kind of all predictions made by the algorithm. For our simplified version we can compute accuracy as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

To be more precise, if we denote the Confusion matrix as C and the set of all possible target values as $Dx = [CN, EMCI, LMCI, SMC, AD]$, then we can compute the accuracy as:

$$Accuracy = \frac{\sum_{i \in Dx} C_{i,i}}{\sum_{i \in Dx} \sum_{j \in Dx} C_{i,j}}$$

Accuracy is a good measure when the target variable classes in the data are nearly balanced. It should never be used as a measure when the target variable classes in the data are a majority of one class.

C. Precision

Precision is a measure that tells us what proportion of patients that we diagnosed with one of the diagnosis, actually had that diagnosis.

We calculate precision for every class separately by dividing the number of correct guessed samples of the class and the sum of all samples in the same row. For our simplified version we can compute precision as:

$$Precision = \frac{TP}{TP + FP}$$

In our multi-class model, precision of a particular class is computed as:

$$Precision = \frac{C_{x,x}}{\sum_{i \in Dx} C_{x,i}}$$

where *x* is the label of the particular class.

D. Recall or Sensitivity

Recall is a measure that tells us what proportion of patients that actually had particular diagnosis was diagnosed by the algorithm as having that diagnosis. We calculate recall for every class separately by dividing the number of correct guessed samples of the class and the sum of all samples in the same column. For our simplified version we can compute recall as:

$$Recall = \frac{TP}{TP + FN}$$

In our multi-class model, recall of a particular class is computed as:

$$Recall = \frac{C_{x,x}}{\sum_{i \in Dx} C_{i,x}}$$

where *x* is the label of the particular class.

For example, out of 100 people, 5 people actually have AD. Let's say that the model predicts every case as AD. So, our denominator is 5 and the numerator, person having AD and the model predicting his case as AD is also 5 (since we predicted 5 AD cases correctly). So in this example. we can say that the **recall** of such model is 100% and **precision** of such model is 5%.

It is clear that recall gives us information about a classifier's performance with respect to how many did we miss, while precision gives us information about its performance with respect to how many did we caught. So basically if we want to focus more on **minimising False Negatives**, we would want our **Recall** to be as close to 100% as possible without precision being too bad and if we want to focus on **minimising False Positives**, then our focus should be to make **Precision** as close to 100% as possible.

E. Specificity

Specificity measures a model's ability to correctly generate a negative result for people who don't have the condition that's being tested for. A high-specificity model will correctly rule out almost everyone who does not have the disease and won't generate many false-positive results. A model with a high sensitivity but low specificity results in many patients who are disease free being told of the possibility that they have the disease and are then subject to further investigation.

For example, a model with 90% specificity will correctly return a negative result for 90% of people who don't have the disease, but will return a positive result — a false-positive — for 10% of the people who don't have the disease and should have resulted as negative. For our simplified version, we can compute specificity as:

$$Specificity = \frac{TN}{TN + FP}$$

In our multi-class model, specificity of a particular class is computed as:

$$Specificity = \frac{\sum_{i \neq x} \sum_{j \neq x} C_{i,j}}{\sum_{i \neq x} \sum_{j \neq x} C_{i,j} + \sum_{i \neq x} C_{x,i}}$$

where *x* is the label of the particular class.

F. F1 Score

We don't really want to carry both Precision and Recall in our pockets every time we make a model for solving a classification problem. So it's best if we can get a single score that kind of represents both Precision (P) and Recall (R). One way to do this is simply taking their **arithmetic mean**:

$$F_1 = \frac{P + R}{2}$$

But that's pretty bad in some situations. We need something more balanced than the arithmetic mean and that is **harmonic mean**:

$$F_1 = 2\frac{PR}{P + R}$$

The score lies in the range [0,1] with 1 being ideal and 0 being the worst. Harmonic mean is kind of an average when x and y are equal. But when x and y are different, then it's closer to the smaller number as compared to the larger number. So if one number is really small between precision and recall, the F1 Score kind of raises a flag and is more closer to the smaller number than the bigger one, giving the model an appropriate score rather than just an arithmetic mean.

XV. Evaluation of the RF Classifier

We finished the process of tuning model's hyperparameters and splitting our dataset. Now, it is time to evaluate it's performances. After training our Random Forest model and testing it on the testing data subset, we create Confusion matrix with the predicted labels versus actual labels. On the main diagonal of the matrix are numbers of correct guessed labels from each category respectively.

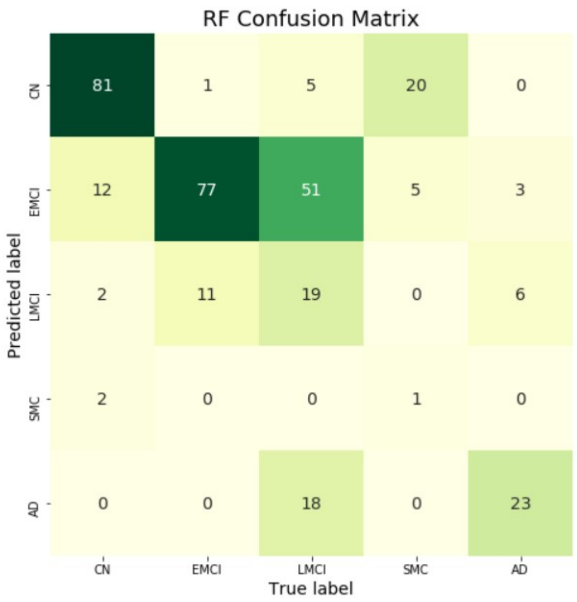
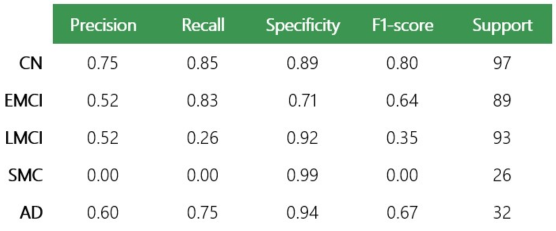


Fig. 24. Confusion matrix for the Random Forest model

As we can see from the figure above, our RF model does pretty good job classifying subjects with CN, EMCI and AD diagnosis. The main problem occurs for subjects with *Significant Memory Concern* diagnosis. It seems like our model is not able to classify any subject with this type of diagnosis correctly. Instead, it classifies these patients as *Cognitive Normal*. It represents big gap in our model, making it unreliable. Misleading information also occurs for subjects diagnosed with LMCI. Our model classify these subjects with EMCI diagnosis.

This phenomenon is somewhat understandable considering the fact that EMCI (*Early Mild Cognitive Impairment*) and LMCI (*Late Mild Cognitive Impairment*) represent two different phases of the same diagnosis. Using the Confusion matrix i.e. values in it's fields, we can compute values of other metrics explained before.

First, we are computing the **accuracy** of the model. The accuracy has a value equal to **0.6**. This is not an excellent result assuming that a subject has around 60% chance of getting his right diagnosis. This mild value is a result of the problem with SMC and LMCI subjects. If we omit them from the accuracy formula, we would get a value of around **0.84**, which is almost excellent accuracy. Unfortunately, this is not how things work in real life.



	Precision	Recall	Specificity	F1-score	Support
CN	0.75	0.85	0.89	0.80	97
EMCI	0.52	0.83	0.71	0.64	89
LMCI	0.52	0.26	0.92	0.35	93
SMC	0.00	0.00	0.99	0.00	26
AD	0.60	0.75	0.94	0.67	32

Fig. 25. Classification report for the RF model

From the previous table we can notice that 26 of the test subjects have SMC diagnosis. In the whole dataset there are 83 subjects with this diagnosis. It means that our model learned about SMC using 57 subjects only. Perhaps this is the problem we are facing and it could potentially be solved if we add more SMC diagnosed subjects to the dataset.

Precision and *Recall* metrics for **SMC** are **0's** since the amount of correct guessed subjects with this type of diagnosis is equal to 1. The result is very small number rounded to 0. But, on the bright side we have a *F1-score* of **CN** equal to **0.8**. The closer this score is to 1, the better it is for our model. As we can see, for **EMCI** we have a *Recall* equal to **0.83** and *Precision* equal to **0.52**. It means that our model guesses correctly patients with EMCI pretty well, but occasionally it labels LMCI patients as EMCI too. For the **LMCI** category, we have a poor *Recall* of **0.26**, which is due to the EMCI impact. **AD** category has a *F1-score* of **0.67** which perfectly represents both Precision and Recall metrics, but it doesn't represent a metric value we should brag with.

XVI. Evaluation of the NN Classifier

We have evaluated our RF model before, gathered it's metrics and now it is time to evaluate our second model - the NN Classifier. We are going to follow the same steps as before, so the first step is to obtain the Confusion matrix.

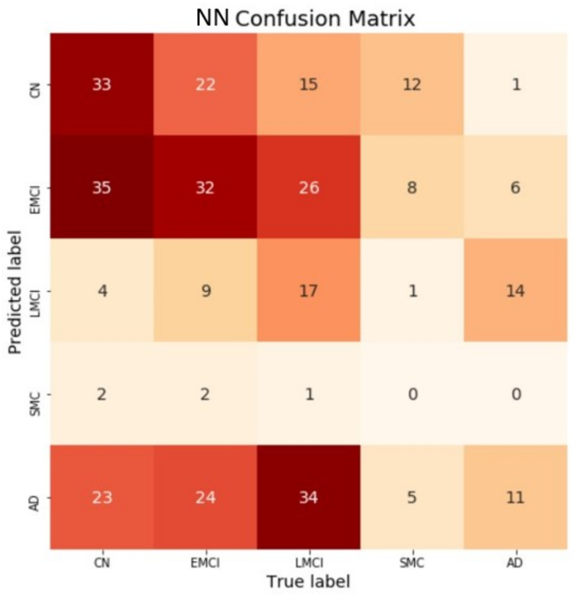
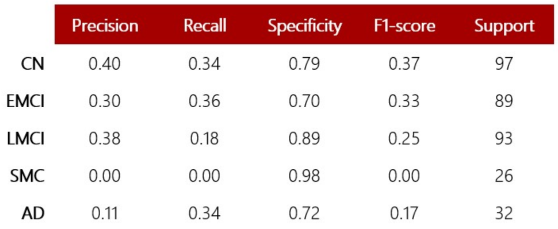


Fig. 26. Confusion matrix for the Neural Network model

This matrix looks a lot more confusing then the previous one. From the very beginning we can assume that our model didn't do very well. Once again, we have total mismatch with the SMC diagnosis. Just like in the previous model, here most of the SMC patients are predicted as CN too. Our model diagnosed most of our LMCI subjects as AD. There is no clear evidence why this mismatch occurred.

Another thing we can notice is the 'dead race' between the CN subjects. 1/3 of the CN subjects are predicted correctly, 1/3 are predicted as EMCI and almost 1/3 are predicted as they have an AD. The last number is little bit confusing since CN and AD are totally contrary diagnoses.

The **accuracy** of this model is equal to **0.276**. This confirms the statement that the NN model is not reliable enough. Observing the process of training and validating, it can be noticed that in every epoch, training and validation **loss values** are very **high** and **accuracies** are pretty **low**. The neural network is not managing to predict even its training results with high accuracy. This is known as **underfitting** and reflects a **low bias** and **low variance** of the model.



	Precision	Recall	Specificity	F1-score	Support
CN	0.40	0.34	0.79	0.37	97
EMCI	0.30	0.36	0.70	0.33	89
LMCI	0.38	0.18	0.89	0.25	93
SMC	0.00	0.00	0.98	0.00	26
AD	0.11	0.34	0.72	0.17	32

Fig. 27. Classification report for the NN model

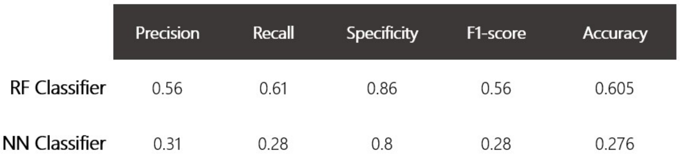
Numbers in the classification report just confirm the statements that were said before. Once again, **SMC** has **0** for each metric. Reason for this is the same as in the RF model.

LMCI has a *Precision* of **0.38**, contrary to the *Recall* that is equal to **0.18**. It means that of all the subjects diagnosed with LMCI, most actually had it. But, most of the LMCI subjects were labeled with wrong diagnosis and this is reflected as low recall. The opposite case is present in the AD class. We have a low *Precision* of **0.11**, but higher *Recall* of **0.34**, meaning that our model classifies subjects with other diagnosis as AD rather than the real AD patients.

Our problem with underfitting gives us untrustworthy results. Couple of experiments had been made with different architectures of the neural network and none of them gave significant improvement. Underfitting can occur for various reasons. One of them is **lack of data**. Perhaps our model doesn't have enough data to learn all dependencies in the data set. Adding more training samples, or improving their quality could potentially solve the problem.

XVII. Discussion

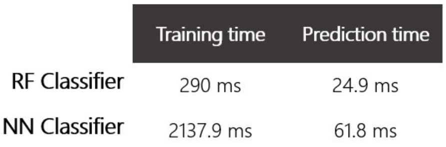
To get a better overview of which model has done a better job, we can summarize the results in one table. For summarizing the categorical metrics, we use a **weighted average** representation. A metric for each label is computed and the average considering the proportion for each label in the dataset is found. In addition, we compare times required to train the models, as well as times required to predict the targets of the testing data set.



	Precision	Recall	Specificity	F1-score	Accuracy
RF Classifier	0.56	0.61	0.86	0.56	0.605
NN Classifier	0.31	0.28	0.8	0.28	0.276

Fig. 28. Comparison of RF and NN model metrics

The Random Forest classifier has a higher value for each metric, indicating that it is more reliable and accurate than the Neural Network classifier. For example, the RF classifier has double the precision of NN. The RF accuracy is almost three times larger than the NN's.



	Training time	Prediction time
RF Classifier	290 ms	24.9 ms
NN Classifier	2137.9 ms	61.8 ms

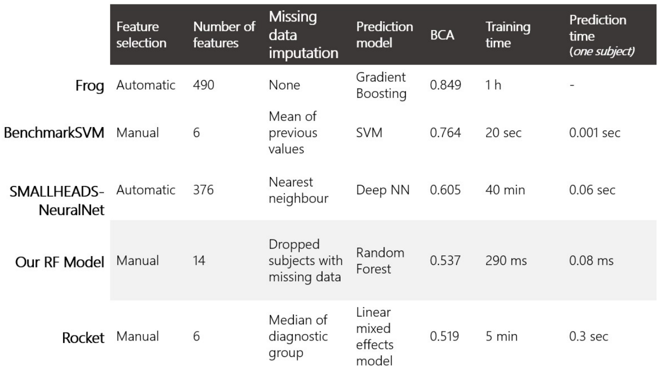
Fig. 29. Comparison of RF and NN execution times

Fig. 8 shows the times required for the models to be trained and to predict the entire test set. Random Forest classifier requires around **seven times less** time to be trained.

Of course, this doesn't make much of a difference, given that the neural network takes about 2 seconds to complete the training process. If a larger dataset was used, the situation would be more critical.

These figures tell us that to solve our problem, **Random Forest is a more optimal algorithm**. The reason for this is probably the dataset we use. Neural networks require a larger and more complex dataset. They are quite good at overcoming complicated problems, but require a large number of samples to learn the dependencies between the features.

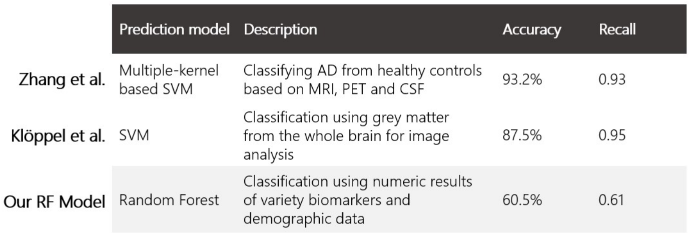
If we compare our Random Forest model with some other models trained on the same original dataset, we can see that our model ranks somewhere below average. The main factor that affects this result is the number of samples that we and other models use. In fact, we removed all the samples where some data was missing, thus significantly reducing the size of our data set. We can see that other models use different types of techniques to fill the gaps.



	Feature selection	Number of features	Missing data imputation	Prediction model	BCA	Training time	Prediction time *(one subject)*
Frog	Automatic	490	None	Gradient Boosting	0.849	1 h	-
BenchmarkSVM	Manual	6	Mean of previous values	SVM	0.764	20 sec	0.001 sec
SMALLHEADS-NeuralNet	Automatic	376	Nearest neighbour	Deep NN	0.605	40 min	0.06 sec
Our RF Model	Manual	14	Dropped subjects with missing data	Random Forest	0.537	290 ms	0.08 ms
Rocket	Manual	6	Median of diagnostic group	Linear mixed effects model	0.519	5 min	0.3 sec

Fig. 30. Comparison of our model with some of the contestants of TADPOLE Challenge

We can also compare our model with the performance of the models processed in the scientific papers mentioned earlier. As we can see, Zhang et al. in their scientific work did a great job distinguishing AD patients from healthy ones using MRI, PET and CSF biomarkers.



	Prediction model	Description	Accuracy	Recall
Zhang et al.	Multiple-kernel based SVM	Classifying AD from healthy controls based on MRI, PET and CSF	93.2%	0.93
Klöppel et al.	SVM	Classification using grey matter from the whole brain for image analysis	87.5%	0.95
Our RF Model	Random Forest	Classification using numeric results of variety biomarkers and demographic data	60.5%	0.61

Fig. 31. Comparison of our model with some esteemed scientific researches.

XVIII. Conclusion

This study proposes classification comparison between two different algorithms. The winner is classification method based on decision trees combination for predicting the right diagnosis for patients among five different diagnoses. This model has some drawbacks, but it is an excellent basis for upgrading to achieve even better and more accurate results. The current study only considers the baseline data of the subjects in ADNI. The main problem for our weak performances was the lack of subjects with complete data. To overcome the limitation of the possible small number of subjects available for training and testing classifier as discussed earlier, more advanced methods in machine learning which can use missing data for classification, i.e., semisupervised classification can be used. We expect that, by using more samples (*with both complete and missing modality information*), the semi-supervised method will improve the classification performance further.

Creating an exact and precise model that can reliably predict a patient's diagnosis and even predict the next stage of his condition is of paramount importance in neurology. Many patients will be able to slow down the progression of the disease in this way, and thus reduce the damage that this disease causes to a person, as well as to his environment.

REFERENCES

[1] NCBI Bookshelf, 'Alzheimer's disease: Overview', 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279360/ [Accessed: 04.06.2020]

[2] Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E.,Alzheimer's disease. Lancet 2011; 377(9770): 1019-1031.

[3] Mayo Clinic, 'Alzheimer's disease - Symptoms and causes', 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447 [Accessed:04.06.2020]

[4] Schaill et al., 'Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI', 2002, Available: https://www.pnas.org/content/99/7/4703

[5] Yang et al., 'Quantifying the Pathophysiological Timeline of Alzheimer's Disease', 2011, Available: https://content.iospress.com/articles/journal-of-alzheimers-disease/jad110551

[6] Doody et al., 'Predicting progression of Alzheimer's disease', 2010, Available:https://alzres.biomedcentral.com/articles/10.1186/alzrt25

[7] Bateman et al., 'Instantiated mixed effects modeling of Alzheimer's disease markers', 2016, Available: https://www.sciencedirect.com/science/article/abs/pii/S1053811916302981

[8] Klöppel et al., 'Automatic Classification of MR Scans in Alzheimer's Disease', 2008, Available: https://pubmed.ncbi.nlm.nih.gov/18202106/

[9] Zhang et al., 'Multimodal classification of Alzheimer's disease and mild cognitive impairment', 2011, Available: https://www.sciencedirect.com/science/article/abs/pii/S1053811911000267

[10] Fonteijn et al., 'An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease', 2012, Available: https://www.sciencedirect.com/science/article/abs/pii/S1053811912000791

[11] Young et al., 'A data-driven model of biomarker changes in sporadic Alzheimer's disease ', 2014, Available: https://academic.oup.com/brain/article/137/9/2564/2848155

[12] Jedynak et al., 'A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort', 2012, Available:https://www.sciencedirect.com/science/article/abs/pii/S1053811912007896

[13] Donohue et al., 'Estimating long-term multivariate progression from short-term data', 2014, Available: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2013.10.003

[14] Durrleman et al., 'Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data', 2013, Available: https://link.springer.com/article/10.1007/s11263-012-0592-x

[15] Lorenzi et al., 'Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images', 2015, Available: https://www.sciencedirect.com/science/article/abs/pii/S0197458014005594?via