UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Yevheniia Kryvenko

Using Machine Learning to Explore Genotype Effects on Cortical Thickness of Human Brain

Master's Thesis (30 ECTS)

Supervisors: Raul Vicente Zafra, Ph.D.

Tartu 2020

Using Machine Learning to Explore Genotype Effects on Cortical Thickness of Human Brain

Abstract: The human brain is one of the most complex and unstudied parts of our body. One way to explore the cerebral cortex is to receive magnetic resonance imaging output and to calculate different measurements like cortical thickness, cortical volume, white surface total area, etc. A researcher might compare the obtained values across the defined population or through historical changes of one particular subject. Since many factors might have an impact on the brain (genetic factors, inheritance, environmental impact, lifestyle, nutrition, education) there exist limitations in the analysis. In this thesis, we aim to examine several chosen genotypes and cortical thickness in many regions of interest across the brain to understand the hidden relationship between them and possible use in early diagnostics. Since neurodegenerative diseases are not easy to diagnose in time, the preventive analysis should be introduced. For example, some genes markers (E4 allele of the APOE gene) are already known to be associated with higher chances of getting Alzheimer's disease and people in high-risk group care more about regular health check-ups. Using machine learning techniques to examine genotype effects on cortical thickness brought some meaningful outcomes for further discussion.

Keywords: statistical testing, classification, cortical thickness, genotypes, categorical regression

CERCS: P170 Computer science, numerical analysis

Masinõpe kasutamine uurimaks genotüübi mõju inimese ajukoore tihedusele

Lühikokkuvõte: Inimese aju on üks meie keerulisemaid ja vähem mõistetavamaid kehaosi. Üks võimalus uurida suurajukoort on kasutada magnetresonantstomograafiat ja arvutada erinevaid suurusi nagu näiteks ajukoore paksust, ajukoore mahutavust, valgeolluse üldpindala, jne. Teadur võib võrrelda saadud väärtusi arvestades määratletud rahvastikku või läbi ajalooliste muutuste ühel konkreetsel subjektil. Kuna inimaju võivad mõjutada mitmed tegurid (geneetilised tegurid, pärilikkus, keskkonna mõju, elustiil, toitumine, haridus), siis esinevad analüüsil piirangud. Selle magistritöö eesmärk on uurida mitmeid valitud genotüüpe ja ajukoore paksust erinevates ajupiirkondades, et mõista peidetud seoseid nende ning võimaliku varajase diagnoosimise kasutusvõimaluste vahel. Kuna neurodegeneratiivseid haiguseid ei ole lihtne õigel ajal diagnoosida, siis tuleb juurutada ennetavat analüüsi. Näiteks mõned geeni markerid (E4 alleel APOE geenis) on seotud kõrgema tõenäosusega saada Alzheimeri tõbi ja kõrgeid riskigruppe jälgitakse regulaarsetel meditsiinilistel ülevaatustel rohkem. Masinõpe kasutamine genotüübi mõju uurimisel ajukoore tihedusele tõi esile olulisi tulemusi edasiseks uurimiseks.

Võtmesõnad: statistiline testimine, klassifitseerimine, ajukoore tihedus, genotüüp, kategooriline regressioon

CERCS: P170 Arvutiteadus, arvutusmeetodid

Contents

1 Introduction

Human brain begins forming in prenatal life and due to its complexity it is needed 25-30 years to fully develop. As we age people are able to observe visual changes with their body but brain is out of our optic scope like skin or muscles. However at some time of senescence declines in memory and cognitive abilities could be interpreted as brain consenescence.

One of the main parts of the human brain is the cerebral cortex which is involved in many important cognitive functions comprising sensory, motor and association areas. Cortex is the thin 2-3 mm layer that covers the external part of the cerebrum and values of cortical thickness vary across the life. Besides aging effects, the thinning process could be treated as an indicator of early neurodegenerative processes. It is a known fact that brain aging affects everyone and is inevitable to some extent but this process is not uniform and hit us differently. The identification of a specific pattern of cortical thinning associated with disease stage and the evolution of cognitive impairment would add to the available evidence concerning cortical involvement in neurodegenerative disorders [1].

Also, the structure and organization of several brain areas are highly determined by genetics. In particular, cortical thickness and surface area of different brain areas has been proposed to be partly determined by age and genotypes as was discussed in [1, 2, 4, 5]. Genetic influences is high in the corpus callosum and in early-maturing brain areas like the occipital lobes, while environmental influences has bigger impact in frontal brain areas that have a more protracted maturational time-course [5].

Since the dynamic of brain changes across the adult human lifespan is highly nonlinear there is no simple model to predict the progressive degeneration of the structure of the central or peripheral nervous system. Moreover, due to high variability and complexity of the problem there does not exist a clear statement/formula how to describe brain thinning process by gene expression levels. What is important to mention that in research from [2] only univariate interactions between gene and brain area were studied but in this thesis we will try to add multiple genotypes relation to the model of cortical thickness.

In this thesis we aim to explore the relationship between several genotype expressions and brain areas thickness values. Specifically, we proceed the data collected from 257 patients as simultaneous measurements of genotypes and cortical thickness with applying Machine Learning techniques.

The statistical testing has shown some significant relationship between gene expression levels and cortical thickness of several brain areas, while predicting the whole picture of cortical thickness remains unrevealed. The analysis will be performed in two direction.

We refer to building a regression model (on cortical thickness) from categorical features (genotypes) as the direct problem. And the inverse problem is defined as predicting the expression level of the chosen genotype from the pattern of cortical thickness. The limitations in the analysis were solved using techniques described in section 3.2.4 and 3.2.5.

2 Background

During the whole life human brain changes in subtle but measurable way and modern medicine offers a way to find out brain alteration up to microscopic variation in brain cells and chemistry. Next we briefly describe the main technique to measure cortical thickness in humans as well as some of its factors of influence.

2.1 Measuring cortical thickness

The way for brain studying depends on our type of interest. One approach to explore the structural changes during brain aging is measuring cortical thinning and volume loss.

In order to get needed information we should concentrate on methods with good spatial resolution. One of the best known approaches for structural study is the Magnetic Resonance Imaging (MRI). MRI a non-invasive way to get the detailed image of the organ of interest using a strong magnetic field and radio waves.

Using high-quality MRI data (often $1 \times 1 \times 1$ mm image resolution and good tissue contrast is required), it is comparably straightforward to extract a 3D cortical surface overview from an individual subject's scan [4]. Two types of methods are common:

- a 3D surface deformation with a fixed parameterization, such as a spherical mesh, into the configuration of the cortex.
- the white matter surface identification first as a set of voxels, then imposes a surface triangulation on it, after inflates it to a sphere so that the spherical coordinates can be projected back onto the original surface as a basis for subsequent computations [4].

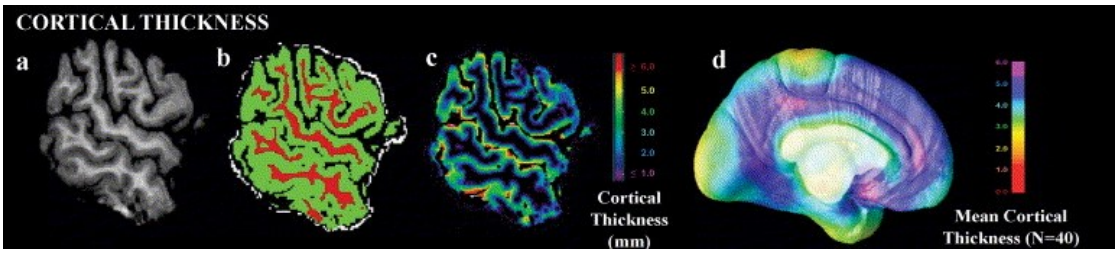


Figure 1. Statistical maps of cortical thickness (a–d): a) the MRI scan b) scan classified into gray matter, white matter, CSF, and a background class by color c) the 3D distance measured. The image was taken from [4].

In the second method (used to obtain the data studied in this thesis), to quantify cortical gray matter thickness, the 3D distance was obtained from the cortical white-gray matter boundary in the tissue-classified brain volumes to the cortical surface (gray–CSF boundary) in each subject (see panel c in Figure 1).

The measurements of gray matter thickness at dozen of homologous cortical regions in each subject is then compared across subjects and averaged at each cortical surface area to get spatially detailed maps (see panel d in Figure 1).

In the cortical surface stream, FreeSurfer (software for processing and analysing brain MRI images) construct models of the boundary between white matter and cortical gray matter as well as the pial surface [7]. Once these surfaces are known, an array of anatomical measures becomes available, including: cortical thickness, surface area, curvature, and surface normal at each point on the cortex.

All together structural measurements give a good overview of our brain organisation and provide numeric values for further statistical analysis.

2.2 Genetic and age impact on cortical thickness

Developmental and adult age-related changes in cortical thickness followed closely the genetic organization of the cerebral cortex, with change rates varying as a function of genetic similarity between regions.

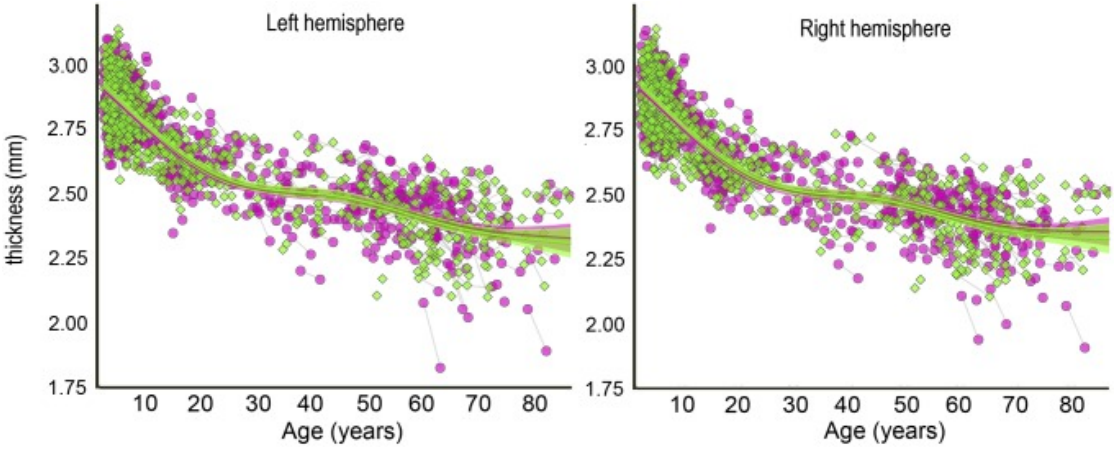


Figure 2. Global change in cortical thickness for each hemisphere in the total sample. Green color refers for female subjects and pink one for male subjects [2]

As discussed in [2], brain areas with overlapping genetic profile demonstrated correlated developmental and adult age affect the trajectories and vice versa for areas with low genetic overlap. Thus, regional cortical thickness variations caused by genes can be traced to local dissimilarity in neurodevelopmental change rates and extrapolated to further adult aging-related cortical thinning. Genetics facilitates cortical changes and raises the question of increase a lifespan perspective in research to improve identifying the genetic and environmental qualifiers of cortical development and aging.

The cortical thinning process continues throughout the remainder of the lifespan (as shown in Figure 2), reflecting reductions in the number of synapses and neuronal shrinkage [2]. And genetic correlations between cortical thickness in the different brain areas can be utilized to parcellate the adult cortex into areas of maximal shared genetic influence. Such an approach can be true according to the belief that genetically programmed neurodevelopmental disorders cause a permanent force on the structure of the cerebral cortex detectable decades later on [2].

3 Methods

In the following section we will introduce the data used for this thesis and applied tools to study the relationship between genotypes and cortical thickness of 148 brain areas.

3.1 Datasets

All analyses were performed using 2 datasets: genotype data and MRI measurements. The data was obtained in collaboration with the Laboratorio de Neurociencia Funcional, Universidad Pablo de Olavide, head by Dr. José Luis Cantero Lorente.

3.1.1 Demographic information

The dataset was collected from 257 subjects and all of them are reported to be healthy. The sex distibution is 131 of female and 126 of male patients. The age statistics are the following: mean value is 66.74 years and standard deviation is 5.33 years.

3.1.2 Genotype dataset

First dataframe contains a description of 6 selected genes for all subjects. Mentioned genes were selected based on statements from several articles which suggest that some of those above mentioned genotypes have an impact on several brain diminution and cognitive impairments. The gene specifications are following:

- ABCA7 - ATP-binding cassette sub-family A member 7 protein.
- APOE - Apolipoprotein E protein.
- BIN1 - Bridging Integrator-1 protein.
- CR1 - Complement receptor type 1 protein.
- CLU - Clusterin protein.
- PICALM - Phosphatidylinositol binding clathrin assembly protein.

3.1.3 MRI measurements

Second data frame contains the measurements of cortical thicknesses at **148 regions** of interest in the brain for each subject separately. The cortical thickness values were extracted from the table of **FreeSurfer** cortical parcellation anatomical statistics per subject. This process involves a series of pre-processing stages of the subjects MRI scans from alignment to tissue segmentation and extraction of cortical anatomical measures is illustrated in Figure 3. The generating program was mris_anatomical_stata.

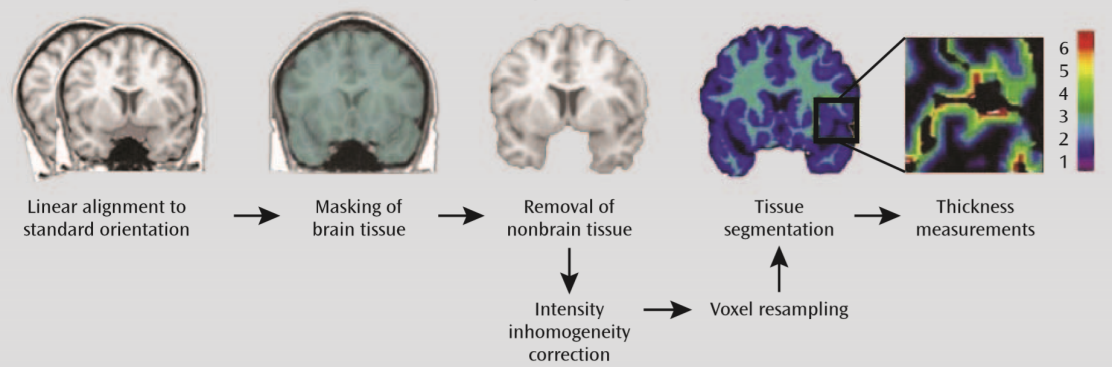


Figure 3. Pre-processing stages of MRI scans to Assess Cortical Thickness: steps required to derive cortical thickness maps from each participant's MR image volume [6].

The FreeSurfer provide opportunity to assign a neuroanatomical label to each location on a cortical surface model based on probabilistic information estimated from a manually labeled training set. The mentioned dataset was created based on the Destrieux atlas parcellation [7], illustrated in the snapshot in Figure 4.

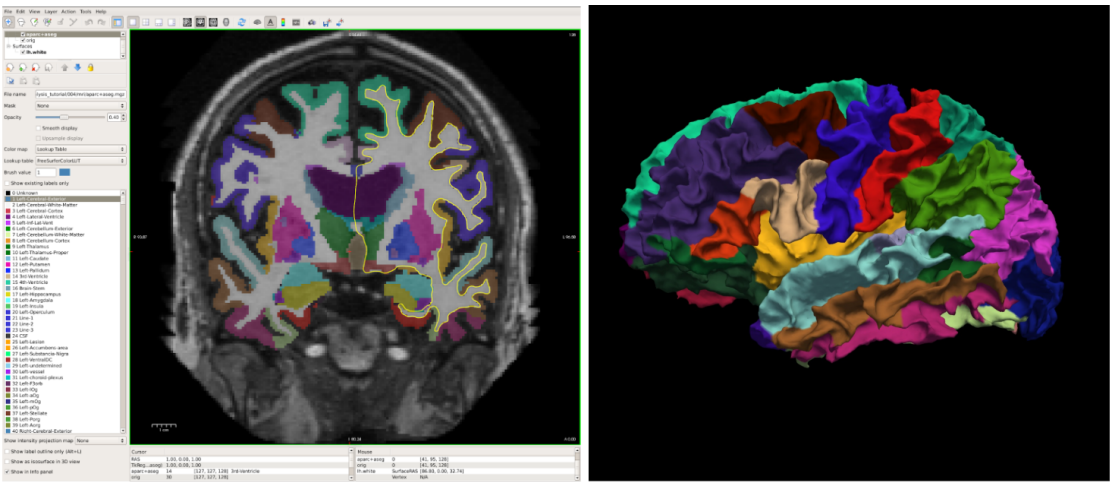


Figure 4. Snapshot of tool with anatomical view of brain ROIs according to Destrieux atlas [8].

Once anatomical ROIs were defined, we are able to extract statistical output from the **cortical parcellation**. The format of the original information is a regular text file and it contains the thickness of specific brain areas. The final numeric values of measurements ready for statistical analysis is represented in format shown in Table 1.

StructName	NumVert	SurfArea	GrayVol	ThickAvg	ThickStd	MeanCurv	GausCurv	FoldInd	CurvInd
G_and_S_frontomargin	1153	803	2093	2.227	0.473	0.144	0.035	17	1.7
G_and_S_occipital_inf	1413	909	2602	2.467	0.695	0.131	0.034	18	1.9
G_and_S_paracentral	1539	883	2347	2.245	0.506	0.103	0.026	14	1.6
G_and_S_subcentral	974	618	1684	2.375	0.415	0.129	0.035	11	1.4
G_and_S_transv_frontopol	837	565	1374	2.135	0.423	0.155	0.048	17	1.7
G_and_S_cingul-Ant	2162	1479	3718	2.326	0.567	0.119	0.03	30	2.7
G_and_S_cingul-Mid-Ant	1441	954	2358	2.37	0.517	0.124	0.027	18	1.5

Table 1. View of small portion of raw input data.

Thus, for each subject and brain area we have structural information including cortical thickness, surface area, gray volume, standard deviation of cortical thickness, curvature, etc. In particular of this thesis we keep focus on cortical thickness of all brain areas for each subject defined as the shortest three-dimensional distance from the cortical white-gray matter boundary to the hemispheric surface without crossing voxels classified as Cerebrospinal fluid (CSF).

3.2 Algorithms

From data description mentioned above we can summarize that for each patient there are 148 continuous variables distributed in approximately the same range and 7 categorical features.

Hence, the main problem addressed in this thesis is that of building a regression model (cortical thickness) from categorical features (genotype). We refer to this as the **direct problem** (from genotype to cortical thickness). Besides univariate random variable prediction we will explore the interaction between genotypes themselves and include the multivariate prediction models.

In addition, we also describe the inverse problem of predicting genotype from the pattern of cortical thickness. Note that while there cannot be a direct causal link, cortical thickness could have predictive value for the genotype of different subjects. We refer to this classification problem as the inverse problem.

To tackle these problems we consider a series of exploratory, regression and classification models that are described next.

3.2.1 Explanatory analysis

Since we do not assume prior knowledge about patients in scope of the project and limited neuroscience background there is an opportunity to gain some knowledge from look into the data. The following steps are conducted to learn about subjects and existing data:

- Discovering structure (correlation) of cortical thickness across different brain areas.
- Analysis of distribution and correlation of genotypes.
- Statistical test of relationship between cortical thickness and genotypes.

The main goal is to check the idea whether it is even possible to explain variability of one variable values based on another one, so we are running the hypothesis testing to select candidates for predictive analysis.

Correlation. Correlation analysis is a statistical procedure to check the linear dependency between two random variables and with this method we do not take into consideration causal association. The output coefficient is a measure of how two variables are related to one another. There exists several types of correlation coefficient. Here we use the Pearson correlation coefficient which measures the linear correlation between two variables X and Y. Values of coefficient varies between -1 and +1, where +/- 1 means total linear correlation, 0 stands for linear independence [9]. The mathematical expression for the Pearson correlation coefficient r_{xy} is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

where:

- n is a sample size.
- x_i, y_i are the individual sample points indexed by i.
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - the sample mean, and analogously for y [9].

Hence, for our purposes a matrix of pairwise correlation coefficients can be used to show the linear dependencies between the cortical thickness of different brain areas.

Software: R studio, function **cor.test()**.

Clustering. Clustering analysis refers to a set of unsupervised techniques for grouping similar instances into the same set. Since this procedure is an iterative process of knowledge discovery it is used for exploratory analysis to get a better understanding of data [10]. Inspecting the outcome produced by hierarchical clustering (dendrogram) is a common way to have an overview of connections/distances at every level/between all instances.

To make a choice on the optimal number of clusters the following auxiliary methods were applied:

- *Elbow method*: suggests a number of clusters based on metric optimizing the within cluster sums of squares.
- *Silhouette method*: suggests a number of clusters based on metric optimizing the average silhouette.
- *Gap statistic*: suggests a number of clusters based on comparison the evidence against the null hypothesis [11].

In our case the clustering analysis can be used to have a visual representation of raw data of cortical thickness measurements, revealing the grouping of brain areas and patients at different levels.

Software: R studio, the functions **hclust(), fviz_nbclust()**.

Principal component analysis (PCA). PCA finds a new set of orthogonal dimensions that are able to capture/represent the maximum of variance across the initial data [12]. Thus, PCA implements a linear transformation that allows to "optimally" project a high-dimensional dataset on a low-dimensional space with new independent variables/axes.

Usual steps to compute the principal components consists of:

- Computing the covariance matrix of the data.
- Calculating the eigenvalues and corresponding eigenvectors of the covariance matrix.
- Normalizing each of the orthogonal eigenvectors to become unit vectors.
- Choosing first k eigenvectors with greatest captured variance.
- Projecting the original n dimensional dataset into new k dimensions.

However, it is a very lucky case to obtain a linear subspace which will be "the best projection" of the initial dataset. Many datasets live on a non-linear manifold or contain non-linear dependencies. Thus, using manifolds learning techniques or kernel PCA we can extend the application of PCA to non-linear manifolds by an explicit construction for data approximation and provide more natural visualization.

The cortical thickness data can be defined as a complex dataset since it consists of 148 features. This issue leads to difficulty with the exploration and visualization of the data. Composing the problem, the combination of high number of features with a low number of instances leads to a severe risk of overfitting and thus worsening the generalization of

prediction models. PCA and kernel PCA can be used for dimensionality reduction and thus to alleviate some of these issues.

Software: R studio, the functions **prcomp(), princomp(), PCA()**.

Chi-square Test of Independence. To check the statement that two categorical variables are associated (i.e. dependent) a Chi-square Test of Independence can be applied to accept or reject the following null hypothesis:

- Null hypothesis: two categorical variables are statistically independent.

The Chi-square test is built in the form of a comparison between the frequencies of each category for one categorical feature across the levels of the second nominal variable. Supposing the resulting p-value is less than 0.05 (for 95% confidence level) we are able to suggest two variables have strong correlation [13].

The chi-square test statistic is calculated as:

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{2}$$

where the expected frequencies e_{ij} are calculated as

$$e_{ij} = \frac{o_i * o_j}{N} \tag{3}$$

and o_i is the marginal frequency of levels of one variable, o_j is the marginal frequency of classes on another variable, and N is the total sample size [13].

During the analysis of genotype dataset Chi-square test is used to examine the associations present between the six categorical features (genes).

Software: R studio, the function **chisq.test()**.

Analysis of variance (ANOVA). In general terms ANOVA is a set of statistical techniques to assess the "variation" among and between groups [14]. The use of ANOVA depends on the research design. Hence, ANOVAs techniques come in different types including one-way ANOVA, two-way ANOVA, and N-way ANOVA. A one-way ANOVA has just one independent variable. Testing one factor at a time hides interactions and can

produce apparently inconsistent experimental results.

Two-way ANOVA testing could be used for two different categorical independent variables on one dependent numerical feature and so on. The ability to detect interactions is a major advantage of multiple factor ANOVA (N-way ANOVA).

During the ANOVA test procedure we check the acceptance or rejection of the following hypothesis:

- Null hypothesis: there is no significant difference among the groups.

The F-statistics is a ratio of two variances and if no meaningful difference discovered between the checked groups, the output of the ANOVA's F-ratio statistic will be close to 1. The formula for F-statistics calculation is following:

$$F = \frac{VarianceBetweenGenotypes}{VarianceWithinGenotypes} = \frac{MS_{genotype}}{MS_{error}} = \frac{SS_{genotype}/(I-1)}{SS_{error}/(n_T - I)} \quad (4)$$

where MS stands for mean square, I is the number of treatments (groups), and n_T is the total number of cases (subjects) [14].

If the probability (p-value) of a value of F greater than or equal to the observed value is smaller than the significance level (in our case p-value < 0.05), then the null hypothesis is considered to be rejected and the alternative hypothesis is supported meaning mean values of all the groups are not equal. The run of post-hoc tests shows which groups are different from each other.

N-ANOVA testing provides insights into the impact of interaction between multiple genotypes on the thickness of one particular brain area.

Software: R studio, the function **aov(), summary.aov()**.

3.2.2 Regression models

Once the statistical modeling is done as a part of explanatory analysis and the causal relationship was revealed with high explanatory power the same techniques or improved one could be used for predictive purposes.

The problem statement is to forecast the cortical thickness value based on known gene's expression for a given patients. Since mentioned task is an example of random

variable prediction the first well-known technique is Linear regression (LR). The linear modeling good to use since it finds a relatively simple model yet explains as much variation as possible.

Categorical regression. Categorical regression refers to the LR model with independent categorical feature. Since the variable with several levels cannot be included directly into a model and be interpreted, additional technique in needed. A categorical variable can be encoded in different ways calling contrast coding system. The easiest method is dummy encoding which simply creates n-1 new dichotomous variables for a categorical feature with n levels of information which keep the same information as a single variable. However, many other kinds of contacts are available like treatment, Helmert, sum and polynomial [15].

After categorical features encoding we are ready to run LR which is similar to ANOVA but now our interest is to obtain the coefficient for predictors (genotypes) and to achieve a good accuracy score. Depending on the number of independent variables there are two approaches for modeling. Single LR fits the relationship between one response variable and one independent variable. In case there is more than one explanatory variable the model is called multiple LR [16].

Linear model evaluates the unknown model parameters from available (training) data thus we are able to get the linear predictive formula to reproduce the relationship.
The formula for single LR is following:

$$y = \beta * X + \epsilon \tag{5}$$

where

- y is a dependent variable
- β - parameter vector / coefficients
- X - input variables / explanatory variables
- ϵ - error term / noise [16].

Software: R studio, the function **lm(), summary(), contr.treatment(), contrasts().**

However the disease can be underpinned by multiple genes contribution those the univariate analysis is not a correct approach usually (we are not able to identify the underlying biological pathways), the multivariate analysis strategy should be applied.

3.2.3 Classifier models

Whenever there is a problem assigning a new observation one of the predefined classes, this is an example of data classification / data labeling task. In general, any class could be unequivocally defined by a collection of mathematical rules.

The classification method complexity depends mainly on data structures. For well-separable data, the rules could be represented as a simple linear function. For linear classification, the most known algorithms are Logistic Regression, the Perceptron algorithm, Support Vector Machine etc.

Classifier gives a solution as a score/probability for all classes per instance and assigns the most likely class. The score is a result of combining a feature vector of an instance with a calculated weights vector. In terms of ML classification is an example of supervised learning which consolidates known input variables in order to assign the predefined classes.

However the large number of features in our dataset complicates the classification rules and arises challenges for the algorithm known as Curse of dimensionality. Hence, as linear rules simply are able to handle high-dimensional data and an ensemble learning method was proposed for such a case.

Decision Tree. The Decision Tree (DT) algorithm is a linear classifier known as an example of supervised learning approach. On each step/node we split instances into two or more homogeneous groups therefore DT usually grows fast and causing overfitting intensely. The method captures irregular patterns easily but has a low bias which leads to high variance.

Ideally, we tend to construct an algorithm with low variance and low bias to elaborate on the prediction which matches correctly following the input data.

Random Forest. Forasmuch the prediction of one single DT may vary greatly the aggregating approach needs to solve this issue. Random Forest (RF) algorithm assemble DTs and assigns the class to an instance based on majority voting of predictions from all trees (see Figure 5). RF approach achieves better results since during the training process each single DT is designed using a set of different picked features and mitigate risks to be skewed by anomaly variables. To sum up, RF arises as to the principal approach for classification in high dimensional space.

In this thesis the RF implementation was taken from scikit-learn library for python.

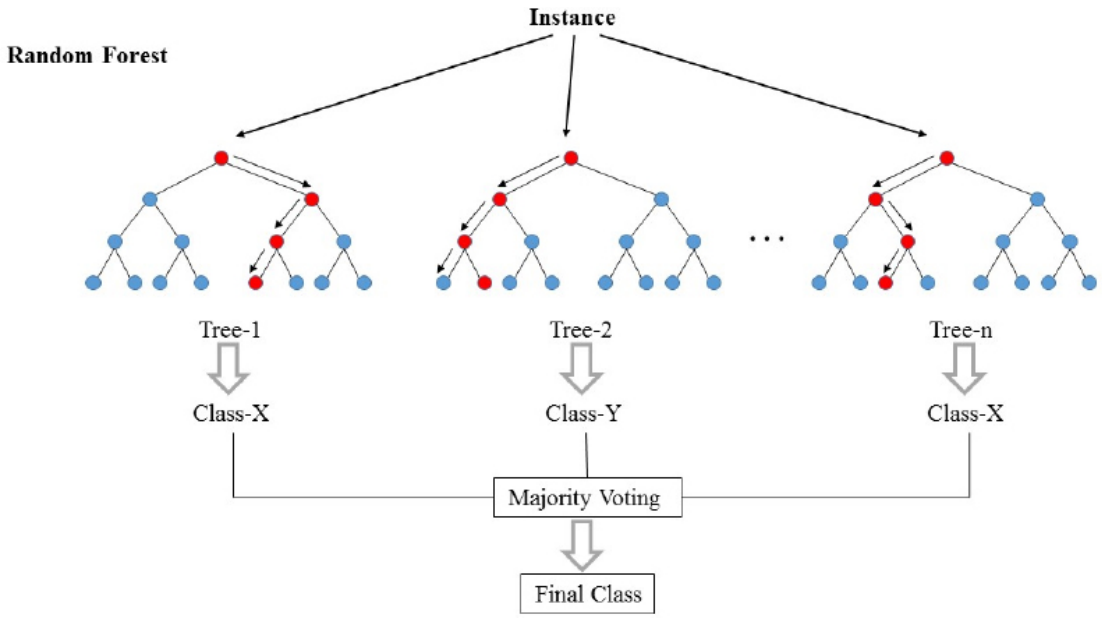


Figure 5. Illustration of RF classification algorithm taken from [28].

3.2.4 Multiple comparison corrections

During the explanatory analysis, we run lots of statistical test iterations and occasionally p-value might be less than a significant level even if the null hypothesis is actually true.

The most common method to manage the error rate (FWER) is the Bonferroni correction which finds the new significant value for an individual test dividing the significant value by the number of tests [17]. However, the Bonferroni correction converts too conservative with a large number of examinations and fails to reject the null hypothesis when it is needed [18].

FWER restraint uses a more powerful control over type I errors in null hypothesis testing compared to the false discovery rate (FDR) procedures. FDR is a method to control the expected proportion of "discoveries" that are false (the number of false positives in all of the rejected hypotheses). This method scans a low proportion of false positives, rather than securing making any false positive conclusion at all. The result is normally improved statistical power and fewer type I errors [19]. In our case with a large number of tests from small samples, we should check q-value. While a p-value of 5% means that 5% of all tests will result in false positives, a q-value of 5% means that 5% of significant results will be false positives [19].

3.2.5 Balancing the classes of data through resampling

The problem of imbalanced class distribution could lead to wrong conclusions on the classification task. ML algorithms working with the dataset where the number of entities belonging to one class is significantly lower than those belonging to another one could be biased and inaccurate. To solve this issue we need to tweak the dataset in order to balance class distribution (balancing classes in the training data - data preprocessing - before input to ML algorithm) and apply ensemble learning technique. Instead of achieving higher overall accuracy we want to improve the identification of rare minority class.

Approaches for partial or full rebalancing the dataset:

- *Oversampling:* increasing the frequency of the minority class, exact replicating some points from the minority class.
- *Undersampling:* decreasing the frequency of the majority class, sampling from majority class.
- *Generating synthetic points:* subsetting the minority class and creating the new similar instances.

4 Results

4.1 Explanatory analysis

Before going to any kind of prediction there should be done an explanatory analysis. Some visualisation, statistical testing, etc. are first steps in any data analysis process and help the researcher to gain some knowledge about working materials.

The main goal is to check the idea whether it is even possible to explain variability of one variable values based on another one, so we are running the hypothesis testing to select candidates for predictive analysis.

We do not assume any prior knowledge about grouping of brain areas or patients and aim to discover a relationship within/between the datasets, hence some unsupervised learning techniques such as clustering and latent variable analysis are considered. In particular, for the explored datasets the explanatory analysis includes:

- Analysis of distribution and correlation of genotypes
  - Genotype data distribution
  - Genes pairwise correlation
- Discovering structure (correlation) of cortical thickness across different brain areas
  - Cortical thickness visualization and clustering
  - Pairwise correlation of cortical thickness
  - Variability of cortical thickness with age
- Statistical testing of relationship between cortical thickness and genotypes

4.1.1 Analysis of distribution and correlation of genotypes

Genotype data distribution The gene expression dataset is represented by 6 categorical variables (genes: abca7, apoe, bin1, cr1, clu, picalm) with different number of levels (5 genes have 3 levels, while APOE has 4 different levels). The distribution of each gene across the sample of 257 subjects is shown in Figure 6 below. From the plot above it is easy to note how imbalanced the dataset is, for example *G/G level* at abca7 gene has only *2%* of subjects while *G/T* and *T/T levels* have *24%* and *74%* of subjects respectively. Indeed, none of the 6 genes has an approximated uniform distribution of levels of expression. In order to achieve a stable and high performance for ML models we will apply some data preprocessing like oversampling or undersampling as explained in section 3.2.5 and used in section (references to methods and results)

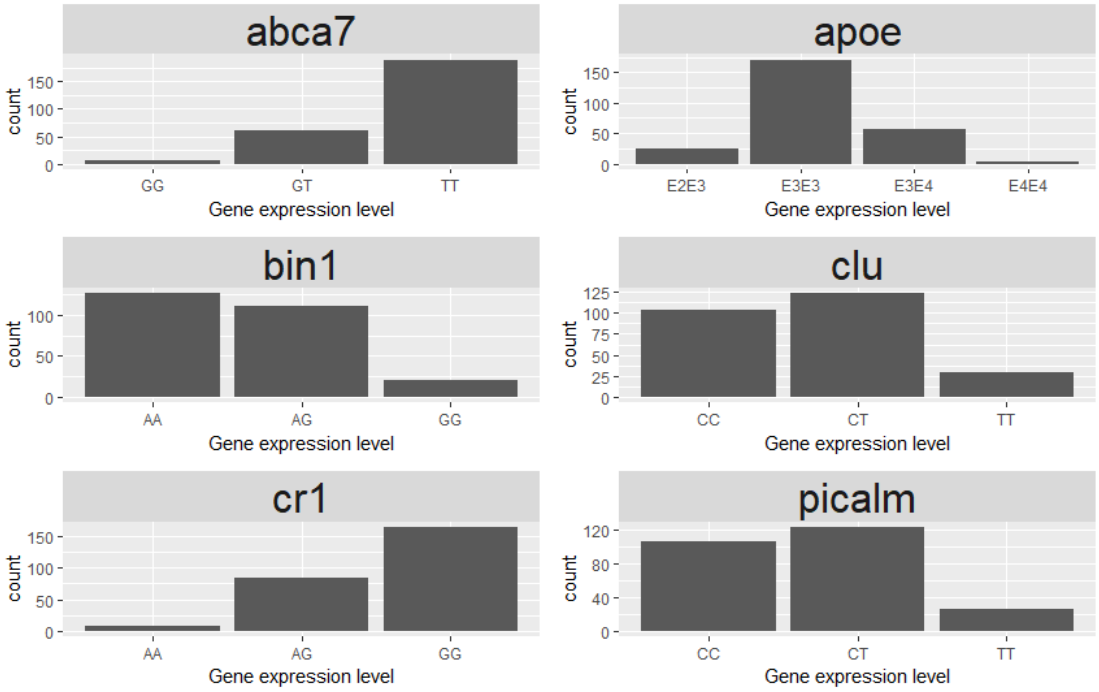


Figure 6. Summary of categorical features in the genotype dataset. The histograms represent the number of counts of each level (expressed genotype) across the sample of 257 subjects.

Pairwise genes correlation Data exploratory analysis requires checking the dependencies within the dataset which is in scope of research. The gene expression data is represented by categorical features (gene alleles) with the additional categorical column as sex. Chi-Squared test was performed pairwise to verify the dependency hypothesis (certain pairs of genes are coexpressed more likely than the chance level according to their marginal distributions).

To sum up there are a total of $\binom{7}{2} = {}^7C_2 = 21$ possible tests which might lead to a multiple comparisons problem. To solve this issue a controlling procedure has to be applied. The most common and straightforward method is the Bonferroni correction however it is known as a very conservative method. The FDR procedures provide less stringent control of Type I errors compared to the previous method, which control the probability of at least one Type I error. Thus, FDR-controlling procedures have greater power, at the cost of increased numbers of Type I errors. Table 2 contains the top 3 pairs of categorical features with lowest p-value after FDR correction.

In particular, one pair of gene-gene coexpression with significant relationship was

Table 2. Chi-Squared test statistics including FDR adjustment. Top 3 pairs of categorical variables according to the FDR corrected p-value.

factor1	factor2	X2stat	pvalue	X2exp	Bonferroni adj	FDR adj
abca7	cr1	17.9946	0.0012	9.4880	0.0260	0.0260
sex	clu	9.1281	0.0104	5.9910	0.2188	0.1094
sex	bin1	6.9556	0.0309	5.9910	0.6484	0.2161

discovered: ABCA7 gene and CR1 gene (see Table 2 for detailed statistics). Since observed p-value is less than 0.05 (for 95% confidence level) we are able to accept the alternative hypothesis and suggest that two variables have strong correlation.

From ALZPEDIA [20] it was noticed that both ABCA7 gene and CR1 gene are involved in susceptibility. Variation in amyloid precursor protein (APP) is identified as a causal factor in Alzheimer's Disease (AD). With respect to AD-related processes, both ABCA7 and CR1 are identified as regulators of APP processing and inhibits Aβ secretion in APP-overexpressed cells [21, 22, 23].

Despite Bonferroni correction being more strictly conservative the amount of significant pairs (p_value < 0.05) after Bonferroni correction and FDR procedure stayed the same for this small number of comparisons.

Table 3. Output of One-way ANOVA tests for age and genotype.

Model	Df	SumSq	MeanSq	Fvalue	Pr(>F)
age vs apoe	3	62	20.51	0.721	0.54
Residuals	253	720	28.46		
age vs bin1	2	100	49.88	1.769	0.173
Residuals	254	7162	28.2		
age vs clu	2	72	36.23	1.28	0.28
Residuals	254	7190	28.31		
age vs abca7	2	27	13.71	0.481	0.619
Residuals	254	7235	28.48		
age vs cr1	2	42	21.13	0.743	0.477
Residuals	254	7220	28.42		
age vs picalm	2	83	41.38	1.464	0.233
Residuals	254	7179	28.26		

Genotype and age Despite the genotype is fixed for each individual we also tested some potential correlation between the age of subjects in the sample and their genotype.

After running the analysis of variance (One-way ANOVA) for each pair genotype and age no significant relationship was revealed (see Table 3 for detailed statistics).

4.1.2 Discovering structure of cortical thickness across different brain areas

Cortical thickness visualization and clustering All subjects are reported to be without any neurological condition and healthy for their age group. Accordingly, all subjects are treated as a single group with no evident stratification besides their age ordering. A possible limitation of the existing data is a lack of information about the patients' background: their education, lifestyle, nutrition, possible head trauma and many other factors that potentially have influence on the brain in total, and cortical thickness specifically. Hence, it might be possible that groups of subjects exist and to conduct a more detailed analysis based on measurements of subjects with similarities within a group. For a possible group detection a clustering procedure was applied to the normalized cortical thickness matrix of subjects x areas.

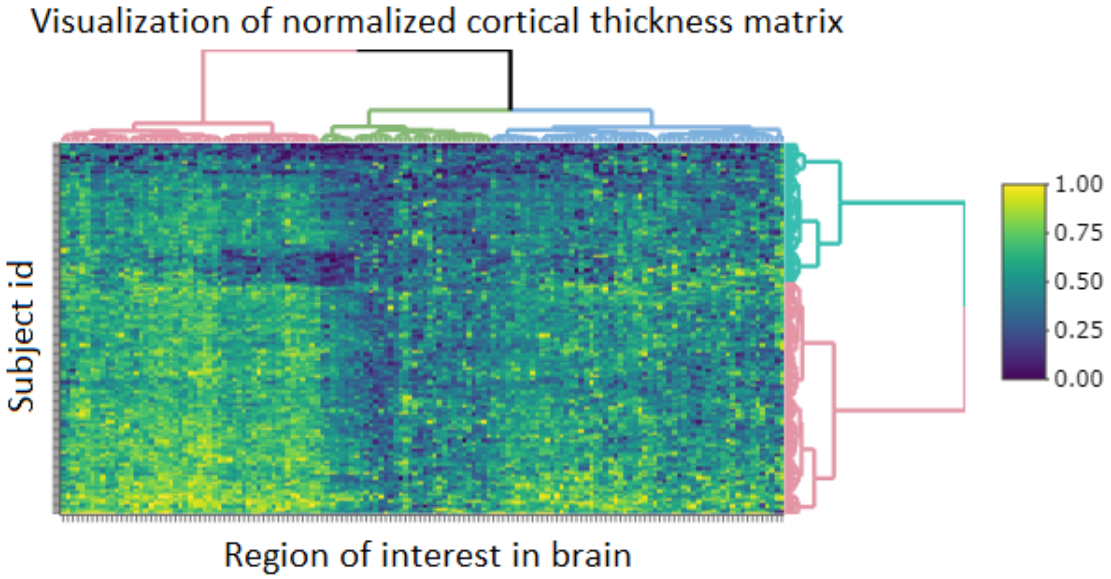


Figure 7. Clustering and dendrograms of subjects and brain areas according to hierarchical clustering on normalized cortical thickness data.

The attempt to organize subjects into clusters and discover brain areas with similar behaviour shown in Figure 7. Clustering was performed using Euclidean distance as a measure of similarity and applied Ward.d2 criteria. According to the clustering analysis of cortical thickness brain areas can be split into 3 groups, while subjects can be grouped into 2 clusters. These findings are consistent across different clustering criteria mentioned in section 3.2.1. The NbClust package in R has fviz_nbclust function to

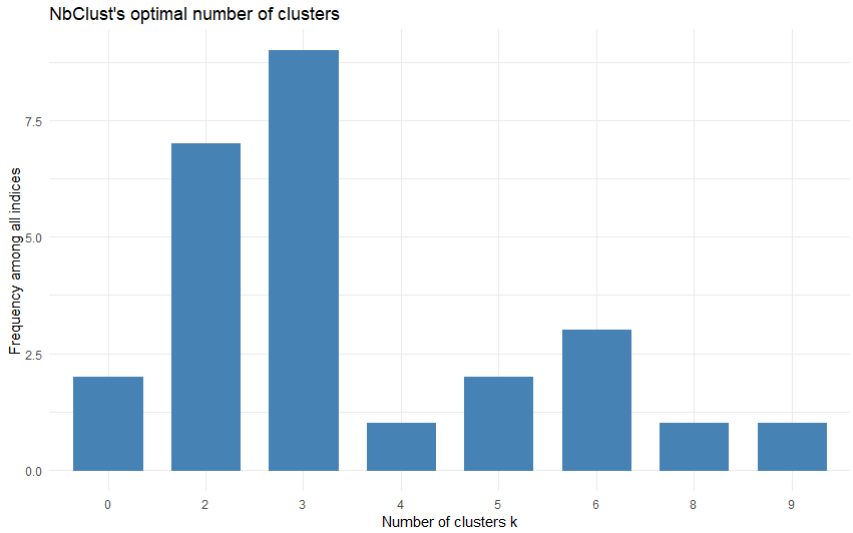
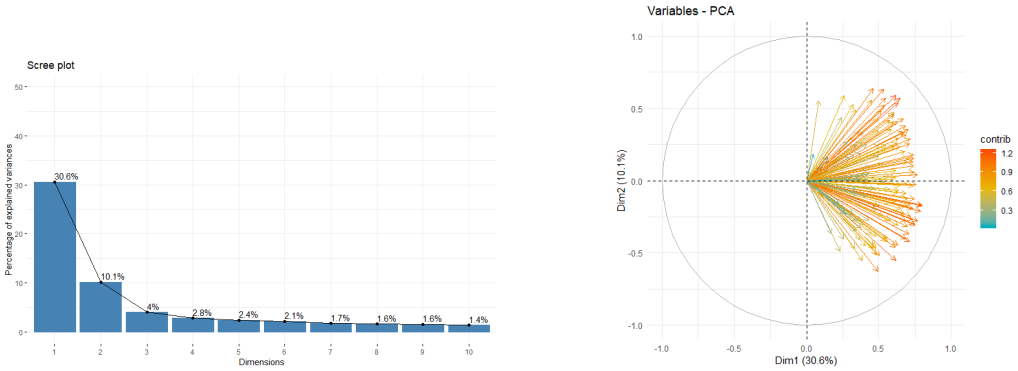


Figure 8. Determination of optimal number of clusters for cortical thickness from all brain areas using 30 different indices.

fit 30 criteria for determining the appropriate number of clusters and suggests the best clustering design from the different outcomes obtained by modifying all combinations of a number of clusters, distance measures, and clustering methods. From Figure 8 the optimal number of clusters is 3 which is aligned with dendrogram on brain areas in Figure 7.

Note that during the working process the data was received in two batches. The results shown in Figure 7 were obtained on the first part of subjects (159 subjects). Based on genotype variables visualisation or sex feature visualisation of any particular cluster, the division on clusters is not connected with genotype nor gender.

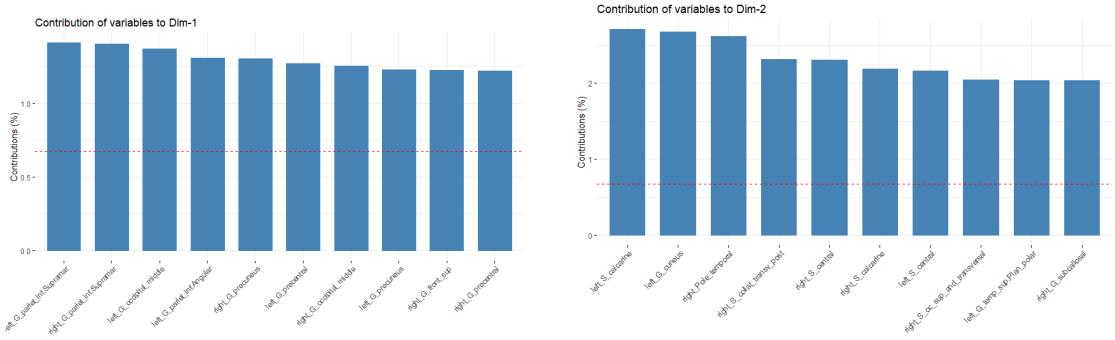
Since there is a large number of dimensions in the data we also experimented with dimensionality reduction techniques to provide some visualization as well as reducing the number of features in predictive models. The first algorithm in scope of the mentioned task was PCA. As seen in Figure 9 (left panel) the first two principal components capture around 40% of the total variance (30.6% and 10.1% respectively) and the rest of the PCs have little contribution. From the visualization of the projected data (Figure 9, right panel), we may conclude that most of the brain areas enhance each other in PC1 (all features are directed in one direction) while in PC2 a balancing mechanism is noticed (features contribute in both directions: positive and negative). Next, we computed how different brain areas contribute to each of the first two principal components. In Figure 10 it is shown the value of the top 10 areas contributing to PC1 and PC2.



(a) PCA scree plot.

(b) Graph of variables

Figure 9. Left panel: Scree plot of first 10 PCs contribution to variance. Right panel: all brain areas contribution and its direction from the cortical thickness dataset to the first two PCs.



(a) Top 10 contributors to PC1.

(b) Top 10 contributors to PC2.

Figure 10. Main brain areas contributors to the PC1 and PC2

Variability of global cortical thickness with age The range of ages for the subjects in scope is relatively narrow (between 53 and 78 y.o.). Previous studies [2] have shown that there is no big change in cortical thickness value after a certain age (around 50 years). For our dataset, the correlation coefficients relating overall cortical thickness and age were as follows:

- Median of cortical thickness of brain areas from left hemisphere vs age: -0.07
- Median of cortical thickness of brain areas from right hemisphere vs age: -0.09

Thus, and given the intersubject variability, there is little variance in global cortical thickness to be explained by the age factor in our dataset.

Pairwise correlation of cortical thickness In order to have an overview of brain areas related to each other based on cortical thickness measurements, the pairwise correlation matrix is plotted in Figure 11. Some brain areas have a strong correlation with a particular set of brain areas while there are brain areas that have almost no correlation at all. Rectangles of the same level of correlations indicate the presence of clusters of brain areas more correlated together. Additionally, there appears to be a strong correlation between the same areas from different hemispheres.

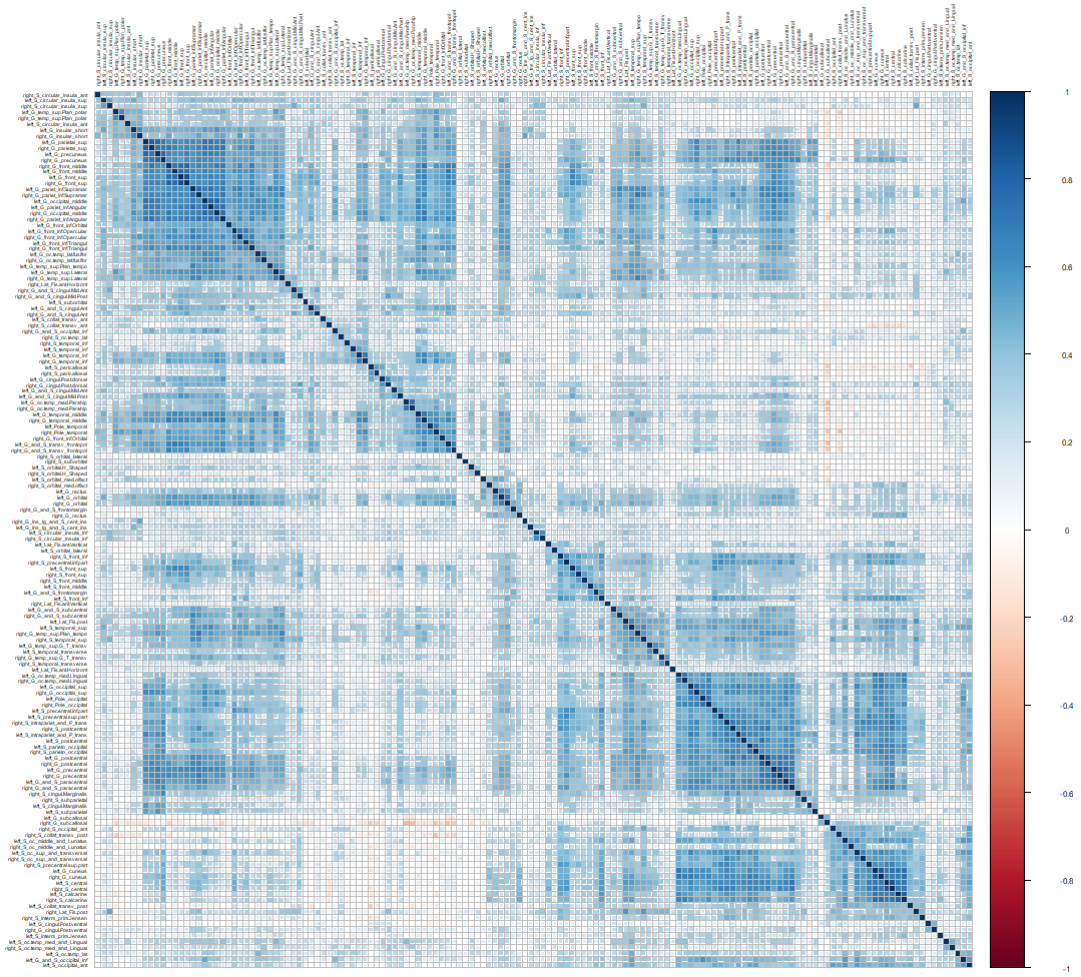


Figure 11. Correlation plot between all ROIs.

4.1.3 Cortical thickness and gene expression interrelation

After exploring each dataset separately now we turn into examining the relationship between cortical thickness of brain areas and genotypes. This will allow us to determine

which variables have an effect on each other and thus are able to explain some part of the variability. The used algorithm is ANOVA test which relies on calculating inter-group variance and intra-group variances to compare them.

We start our analysis with one-way ANOVA test meaning we consider one categorical variable (gene and sex) and one numerical feature (cortical thickness of brain areas) to investigate. Since the dataset consists of 148 ROIs against 6 genes and sex variable there is need to run 1036 iterations of pairwise testing.

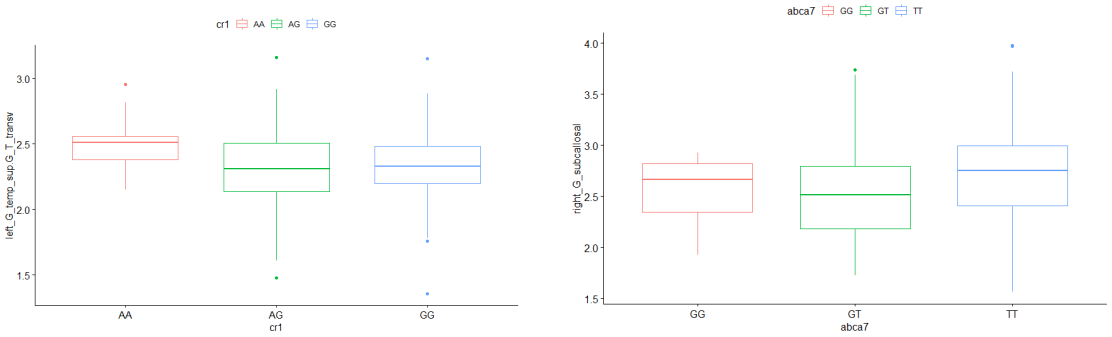
Due to big amount of comparisons some discovered associations could be treated as significant by chance. To avoid such cases we have introduced FDR correction. From one-way ANOVA output the following conjunctions may be interesting to examine closely (see Table 4).

Table 4. One-way ANOVA output: revealed 8 pairs (ROI and sex variable) with significant relationship.

Brain area	**adjusted p-value**
left_G_and_S_cingul.Mid.Ant	0.0092
left_S_collat_transv_post	0.0219
left_S_oc_middle_and_Lunatus	0.0379
left_S_oc.temp_med_and_Lingual	0.0380
left_S_temporal_transverse	0.0219
right_G_and_S_cingul.Mid.Post	0.0379
right_S_oc.temp_med_and_Lingual	0.0002
right_S_temporal_transverse	0.0219

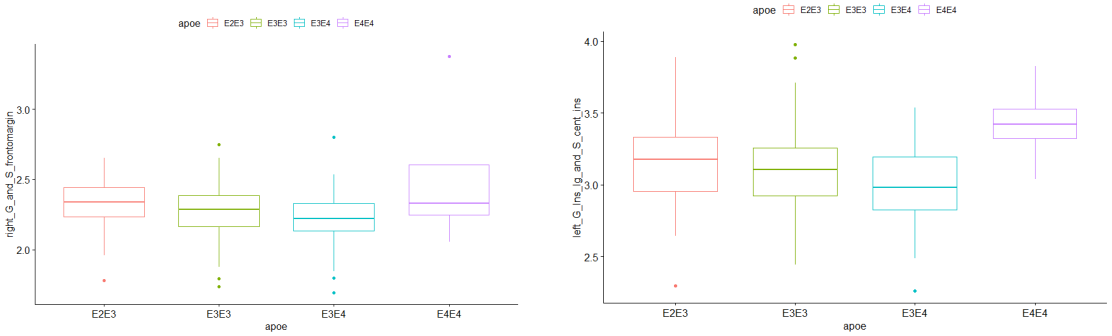
From Table 4 we might notice there is only brain area and sex variable pairs are presented. It means no pairs of brain areas and genotype with the significant relationships were revealed. I agree with the obtained result since from Figure 12 it is noticeable there is no clear separation between the mean value of each group. Mentioned in Figure 12 pairs have the smallest p-value from all single brain area and genotype comparisons. Despite the fact that no significant relationship between cortical thickness in a particular brain area and single genotype were revealed after FDR correction of ANOVA output some interesting combinations for detailed investigation are mentioned later on.

Figure 12 shows that the cohort split is far from being perfect, however it is the best results we have managed to obtain and at least group mean of cortical thickness between the groups differs.



(a) (b)

Figure 12. Distribution of cortical thickness values of specific brain area within genotype (panel a: cr1 gene, panel b: abca7 gene).



(a) (b)

Figure 13. Distribution of cortical thickness values of specific brain area (panel a: right_G_and_S_frontomargin, panel b: left_G_Ins_lg_and_S_cent_ins) within apoe gene.

An important comment to be mentioned is the degree of data imbalance which has an impact on determining the group mean value. The distribution of expression level values across each genotype is highly skewed and for example, the mean of E4 value of APOE group was calculated based on only 4 data points.

From ANOVA test output an interesting example is shown in Figure 13. The known fact is that the E4 allele of the APOE gene is the strongest genetic risk factor for late-onset Alzheimer's Disease and the experiments have proved this gene expression value differs a lot from the rest three possible alleles. APOE E4 is called a risk-factor gene since it increases a person's risk of developing the disease. However, inheriting an APOE E4 allele does not mean that a person will surely develop Alzheimer's. Some people with an

APOE 4 allele never get the disorder and others who exhibit Alzheimer's do not have any APOE 4 alleles [24, 25].

One important remark from all performed test iterations: **No relationship revealed between average cortical thickness and any single genotype.**

The cortical thickness and gene expression relationship could be so complex that pairwise comparison is simply not able to capture/discover it. Thus, there exists the possibility that the effectiveness of one or another gene separately is rather less than when they are both included in the model. To discover the joint effect of predictors we need to run two-way ANOVA test (or three-way ANOVA test) depending on how many genes we include in the analysis.

Table 5. Two-way ANOVA output: combinations of ROI and apoe+bin1 genes with significant relationship.

Brain area	**adjusted p-value**
left_G_subcallosal	0.0412
right_G_and_S_frontomargin	0.0035

Table 6. Two-way ANOVA output: combinations of ROI and apoe+clu genes with significant relationship.

Brain area	**adjusted p-value**
left_G_occipital_sup	0.0048
left_G_pariet_inf.Angular	0.0481
left_S_central	0.0481
left_S_intrapariet_and_P_trans	0.0481
left_S_oc_middle_and_Lunatus	0.0048
right_Lat_Fis.ant.Vertical	0.0365
right_S_intrapariet_and_P_trans	0.0242
right_S_precentral.sup.par	0.0481
right_S_temporal_sup	0.0152

After including the interaction between genotypes to the model the number of significant results increased notably. As a result, 51 combinations of brain area and two genes out of 6216 possible combinations were revealed with a significant relationship. The pair of apoe and bin1 genes has significant relationship with two brain ares (see Table 5), while the pair of apoe and clu genes interact with nine brain areas (see Table 6). The combination of apoe gene and cr1 gene has relationship only with one brain area (see

Table 7) and the rest 39 statistically proven combinations of brain areas and couple of genotypes were discovered for apoe and picalm genotype tandem (see Tables 9, 10 in Appendix).

Table 7. Two-way ANOVA output: combinations of ROI and apoe+cr1 genes with significant relationship.

Brain area	**adjusted p-value**
right_G_oc.temp_med.Lingual	0.0.0181

Results of Two-way ANOVA proved our assumption about the complexity of genotype and brain structure interaction. While there was no meaningful outcome of the One-way ANOVA testing for single genotypes some pairwise genes have a significant relationship with cortical thickness of certain brain areas. However, including more than two genes to the model has not brought more combinations for further predictive part of this thesis. There is only one alliance with a relationship less than the significant level for Three-way ANOVA shown (see Figure 14) but the rest N-way ANOVA models have not worked out.

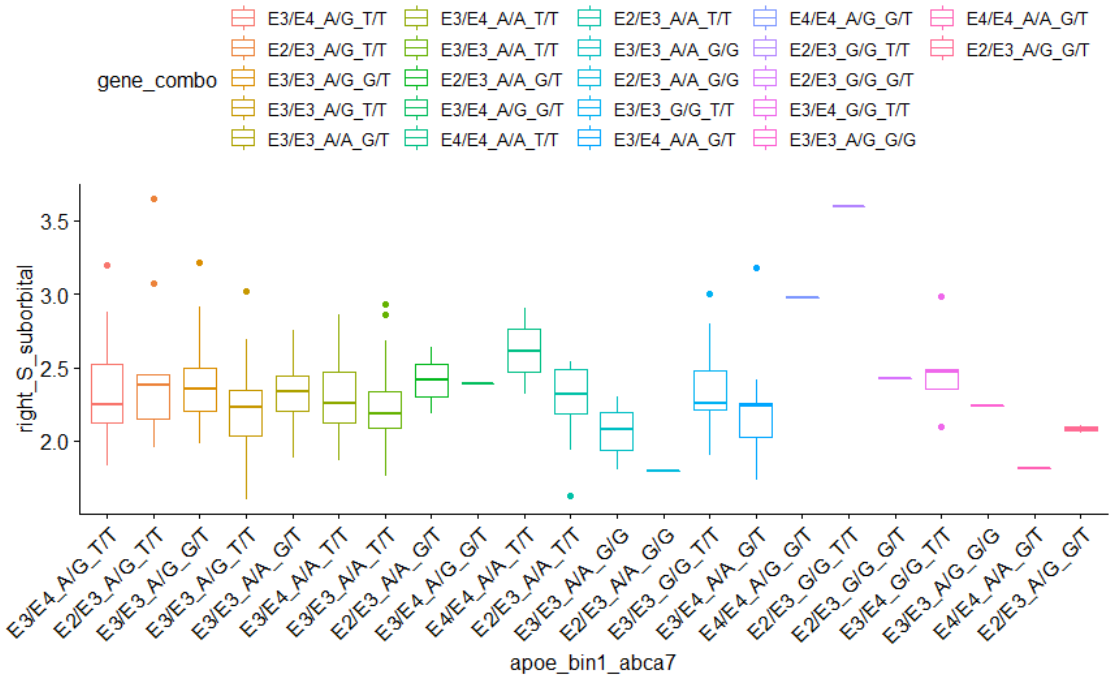


Figure 14. Distribution of cortical thickness values of specific brain area within gene expression levels for combinationn revealed by the Three-way ANOVA.

From Figure 14 we observe the split of cortical thickness of right_S_suborbital brain area across the combination of gene expression levels of apoe, bin1 and abca7. We might agree with the outcome since mean value across all groups seems to be distinguishable from the rest one (although one concern is the limited number of observation for some rare combinations which skew the result). After revealing the significant connections we are able to run the predictive analysis later in this thesis.

4.2 Modelling the relationship between gene expression and cortical thickness

4.2.1 Direct problem: predicting cortical thickness

Nevertheless, the discovery of causal relationship between two datasets at any direction is exciting research problem, we have to focus on empirical prediction as obtaining expression of genes is easier task then measuring patient's cortical thickness which involves expensive equipment and is time consuming process.
The previous part of work helped to filter the alliances with meaningful relationship and in this part of the work we will try to the predictive models. The interest of predictive modeling is an attempt to get cortical thickness prediction from available genotype data. And to measure the effectiveness of generation good predictions for the future outcome the success criteria like predictive accuracy is used.

One issue with categorical regression was noticed and mitigated during experiments that the order in which variables are added to model or removed can affect the final result. Also the order how levels of categorical feature are organized and included to the model has impact on the outcome.

The example shown in Table 8 proves the concept that a single genotype is not as powerful as a combination of a couple of them. The idea that the difference between the two models has a quantity could be measured by the reduction in the unexplained sum of squares produced by the additional variable to the model. To assess the mentioned metric we need to run an F-test, hence apply an ANOVA.

The idea that the difference between the two models has a quantity could be measured by the reduction in the unexplained sum of squares produced by the additional variable to the model. To assess the mentioned metric we need to run an F-test, hence apply an ANOVA. From the ANOVA output shown in p-value = 0.0018 (which is less than the significance level) indicates that the adding bin1 to the model results in a statistically significant decrease in the unexplained variation.

During this part of thesis different other ways of relationship modeling were used.

Table 8. An example of successful discovery: marginal values have no impact on the cortical thickness of the particular brain area while the interaction of genotypes has a positive outcome.

Model	R-Squared	R-adjusted	p-value
left_G_subcallosal vs apoe	0.0032	-0.008	0.8427
left_G_subcallosal vs bin1	0.0006	-0.007	0.926
left_G_subcallosal vs (apoe and bin1 without interaction)	0.0039	-0.015	0.962
left_G_subcallosal vs (apoe and bin1 with interaction)	0.0947	0.0579	0.0055
left_G_subcallosal vs (apoe and bin1 only interaction)	0.0947	0.0579	0.0055

Instead of predicting the cortical thickness of a single brain area we moved to prediction of PCA output on cortical thickness dataset. Since the first 2 PCs captures around 40% of total variability the idea to run LR on projected data seemed to be rational. The results showed that PC1 can't be defined by cortical thickness from any brain areas. However, the p-value of LR between PC2 and age is equal to 3.07e-11 (which indicate the results are significant) and the Adjusted R-squared metric is equal to 0.1636 meaning the model is able to capture 16% of data variability. Although the listed values are far from being perfect, it is pretty high numbers compare to the rest outputs which were close to 0.

4.2.2 Inverse problem: predicting subject genotype

Now we move to another track of research problem and want: Predicting genotype value is a pure example of classification task. However since genotype dataset consists of text labels it is necessary to replace all categorical values with numerical analogs due to limitation of ML algorithms' implementations. For our case the simple ordinal coding is enough and once this step is done we can move to the execution of different ML approaches.

One important remark, the initial gene expression dataset is highly imbalanced. For example, the distribution of coded apoe values is the following:

- Counter 'E2/E3': 161, 'E3/E3':53, 'E3/E4': 25, 'E4/E4':4

To resolve the aforementioned problem the artificial dataset was created using SMOTE and ADASYN algorithms. The first dataset does not contain any duplicate observations (but for some genes we need more samples to solve cases of classes with only 1 or 2 instances). While the SMOTE algorithm uses the nearest neighbors of observations

to create synthetic data, it still bleeds information. If the nearest neighbors of minority class observations in the training set end up in the validation set, their information is partially captured by the synthetic data in the training set.

The accuracy score of Random Forest Classifier (chosen after comparison with other algorithms due to stable performance) on full dataset is low while the misclassification rate of minor classes is very high. To get a clear understanding of algorithm performance F1-score metric was chosen. The low F1-score value helps reveal problems with false positives or false negatives. RF classifier has lots of parameters to tune. For example number of "trees" (estimators) was set to 11 based on cross-validation with values in range [1, 40].

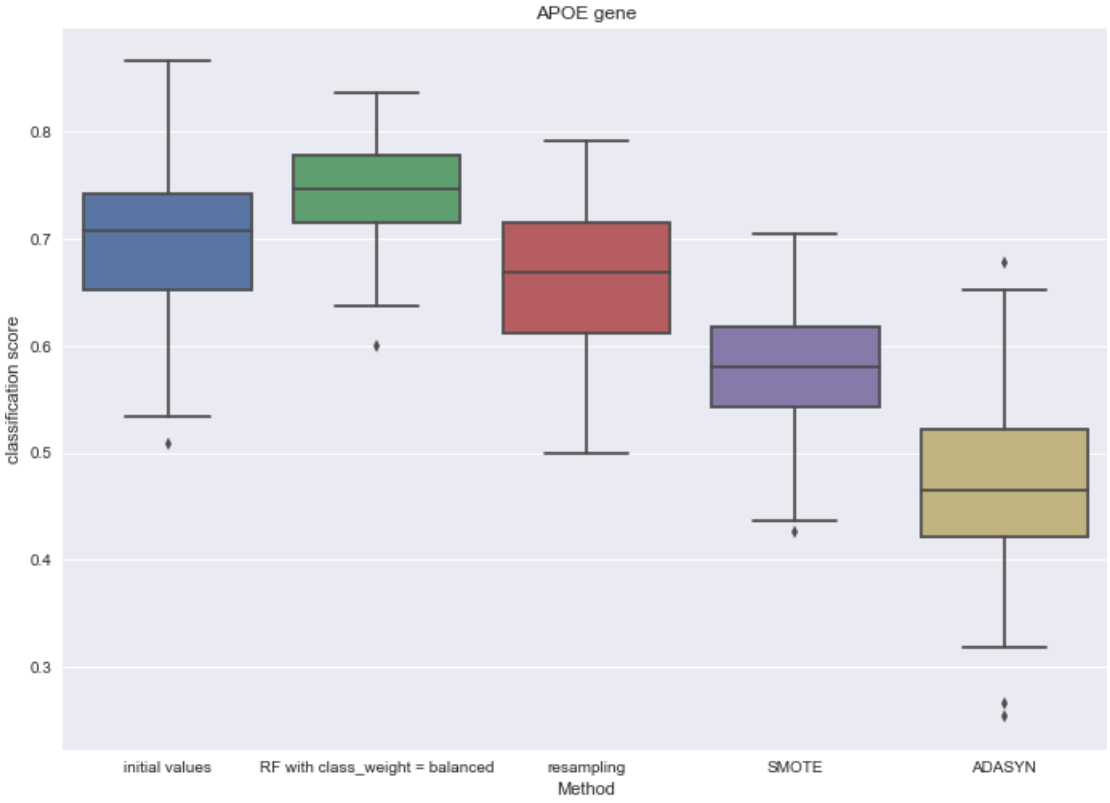


Figure 15. The RF classifier performance on apoe gene prediction from cortical thickness of all brain areas.

The results of 100 iterations of RF classifier on initial dataset, dataset after resampling technique (copying existing minor class instances), SMOTE and ADASYN generated datasets are shown in Figure 15. to compare the performance. During cross validation F1-score metric was used instead of common accuracy score to convey the balance

between precision and recall metrics. After the result analysis it is important to mention that using the resampling method is not correct as the situation of repeating the same instances both in training set and test data could easily happen and we cannot be confident in this classifier on unseen data. The highest score was achieved by RF classifier with automatically adjusted weights inversely proportional to class frequencies in the input data. The next step is to include feature importance and feature engineering in addition to RF classifier and decide what we can do after achievement stable high performance in the classification task.

Feature importance with Gini index On each step of DT building the algorithm select a feature for best data split into smaller dataset to predit the target feature. Each decision is made to maximize the information gain using the impurity criterion (Gini index in our case).

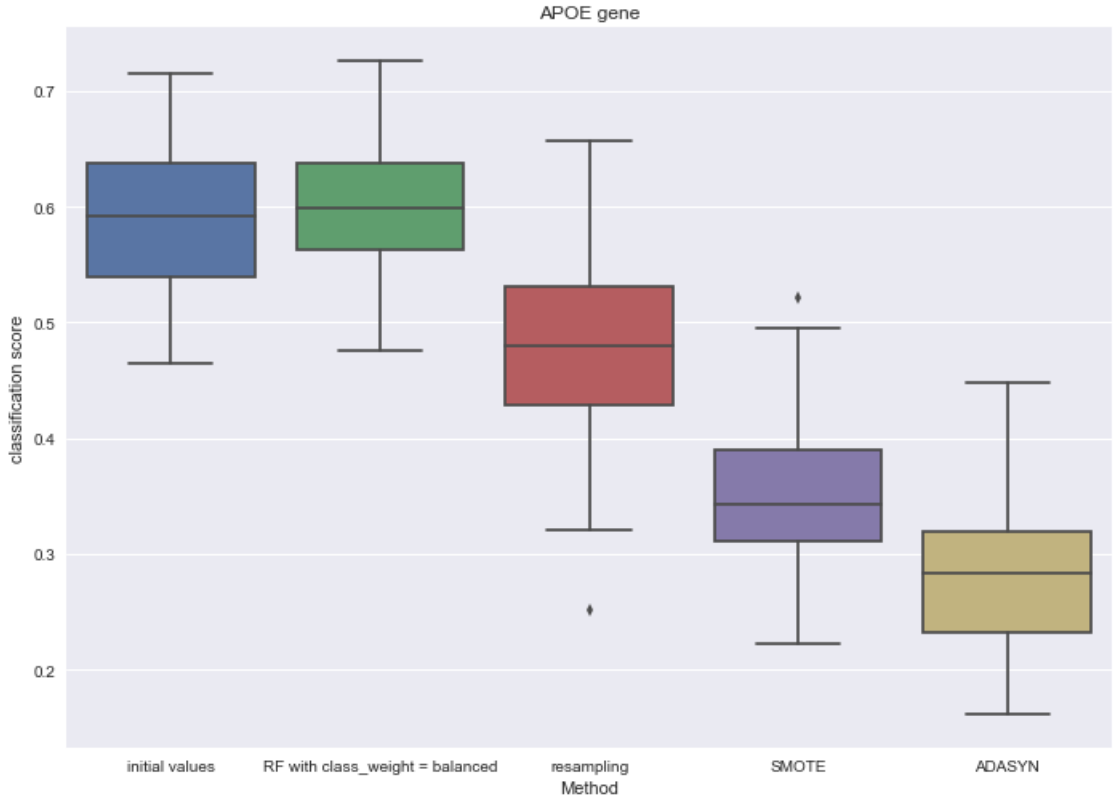


Figure 16. The RF classifier performance on apoe gene prediction from cortical thickness at left_G_front_sup and left_G_subcallosal.

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. Some ROIs have bigger impact than the others. Here, we have extracted feature importance values for each brain area and checked them across all five aforementioned approaches. The following 5 ROIs consistently have a higher feature importance:

- left_G_front_sup
- left_G_subcallosal
- left_G_precuneus
- left_S_occipital_ant
- left_Lat_Fis.ant.Horizont

There still exists a high chance that another brain area will have a big enough impact on the model performance during a random iteration but the mentioned ROIs were picked up due to frequency being in the top of the feature importance list. As shown in Figure 16 the RF classifier is able to perform well even with only 2 ROIs as input instead of a whole set of 148 brain areas which proves that listed above areas effect more on determining the expression level of apoe genotype than any other incidentally chosen area.

5 Discussion

At the beginning of this thesis, we have raised a question: "Do the 6 listed genes (abca7, apoe, bin1, clu, cr1, picalm) determine the cortical thickness in the human brain in the sampled population?". We have tried to figure out the answer using different statistical and machine learning techniques. The outcome of applied approaches showed that none single gene had a statistical relation to cortical thickness. However, we found that models accounting for interactions of pairs of genes (like apoe and bin1 pair, or apoe and clu, or apoe and cr1, or apoe and picalm) were statistically related to cortical thickness of certain brain areas.

As was stated in [3], the possible genetics influence on adult age trajectories could be sought either as a life-long impact on cortical thickness caused by genes through an early development or as an effect of these genes continuously through life and thereby can be detected at widely different ages. Such a statement proves that the relationship between genetics and brain development exists but there are still a lot of challenges to make a clear statement about it.

Meanwhile our findings presented in 4.1.3 are in line with the statement from [2] that there were no significant associations between GWAS dataset and global cortical thickness proved. However at the same time authors have discovered relationship between 16 independent loci across 8 chromosomes determining the cortical thickness value of 9 brain regions while we have stated the existence of the 51 pairs of cortical thickness and gene's expression levels with significant relationship (see Tables 5, 6, 7, 9, 10).

In addition, we also explored the pattern of cortical thickness in itself (not in relation to conditioning by genotypes). PCA analysis revealed that a large fraction of variance of cortical thickness (around 40%) could be explained by first two principal components. Also, most brain areas contribution were aligned with growth of first component but not with second one.

Interestingly, we also found that the inverse problem (predicting genotype variants from the pattern of cortical thickness) was possible for certain genes. While such relationship cannot be causal in the sense that cortical thickness cannot affect or change genotypes, it is likely that both variables (pattern of cortical thickness) and genotype variants can correlate via other variables or common drives not included in the present analysis.

5.1 Limitations

During cortical thickness prediction some limitations of categorical regression were discovered. The results of the models differs depending on encoding order used for independent variables (categorical levels of genes). In addition, since the genotype data is highly imbalanced we need more samples to cover rare values of categorical data and improve the performance. As was discussed in results the pairwise gene interactions brings meaningful outcome of genotype versus cortical thickness analysis therefore including more genes in the scope of research should increase the number of discoveries. In addition, measuring gene expression levels of the involved genes (continuous variable) rather than genotype variants (categorical variable) might help achieve better cortical thickness prediction.

5.2 Future Work

In this thesis many results of cortical thickness and genotype codependency were revealed. But besides presented statistical significance of relationship the neuroscientific interpretation of results is much needed. Domain knowledge expertise of affected brain ares will add value to model calibration and might lead to performance improvement. Also, visualization of the results using brain atlases in 3D space brings better understanding of the problem perception. One of the potential research enhancement is taking into account pattern of genes co-expression and add more genes in scope of analysis. Furthermore, currently from the inverse problem's perspective, the number of instances is close to the number of features, those adding more samples to the already available data will increase the volume of analysis and provide better machine learning techniques efficiency. Also, other available measurements like white surface total area, brain segmentation volume, supratentorial volume could be used to explain genotype and brain structure relationships. And since we have already discussed the patterns of brain areas this information allows us to use more machine learning algorithms might like multivariate regression: predicting pattern of areas instead of a single one.

6 Conclusion

In this thesis we have raised the question of relationship between specific set of genes and cortical thickness based on measurements collected from 257 patients. After reading articles we had a contradictory view. It was stated that some neurodegenerative diseases are associated with certain genes, to be more precise a disease risk factor is described by one specific expression level value of a gene. Also neurodegenerative disorders are associated with depletion of the cortex. But there is a very limited number of positive results are presented in papers about establishing a relationship between the cortical thickness and genotypes.

After all performed analysis and tweaks made we might state there is positive outcome from the investigation of relationship between cortical thickness and expression level of genotypes. However, more advanced non-linear techniques should be applied in collaboration with domain knowledge expertise to bring more meaningful results.

References

[1] Zarei, Mojtaba, et al. *Cortical thinning is associated with disease stages and dementia in Parkinson's disease.* J Neurol Neurosurg Psychiatry 84.8 (2013): 875-882. https://jnnp.bmj.com/content/84/8/875

[2] Fjell, Anders M., et al. *Development and aging of cortical thickness correspond to genetic organization patterns.* Proceedings of the National Academy of Sciences 112.50 (2015): 15462-15467. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687601/

[3] Edith Hofer, et al. *Genetic Determinants of Cortical Structure (Thickness, Surface Area and Volumes) among Disease Free Adults in the CHARGE Consortium.* https://doi.org/10.1101/409649.

[4] Thompson, Paul M., et al. *Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia.* Neuroimage 23 (2004): S2-S18. https://www.sciencedirect.com/science/article/pii/S1053811904003994

[5] Brun, Caroline C., et al.*Mapping the regional influence of genetics on brain structure variability—a tensor-based morphometry study.* Neuroimage 48.1 (2009): 37-49. https://www.sciencedirect.com/science/article/pii/S1053811909004947

[6] Foland-Ross, Lara C., et al. *Investigation of cortical thickness abnormalities in lithium-free adults with bipolar I disorder using cortical pattern matching.* American Journal of Psychiatry 168.5 (2011): 530-539. https://www.researchgate.net/publication/49802057

[7] Free Surfer Analysis Pipeline Overview. In Free Surfer Wiki. https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferAnalysisPipelineOverview

[8] Anatomical Region of interest. In Free Surfer Wiki. https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/AnatomicalROI

[9] Pearson correlation coefficient. In Wikipedia, 31 December 2019 https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[10] Cluster analysis. In Wikipedia, 20 December 2019 https://en.wikipedia.org/wiki/Cluster_analysis

[11] Charrad, M., et al. *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.* Journal of Statistical Software 61 (2014): 1–36. https://www.sciencedirect.com/science/article/pii/S1053811909004947

[12] Principal component analysis. In Wikipedia, 18 December 2019
https://en.wikipedia.org/wiki/Principal_component_analysis

[13] McHugh, Mary L. *The chi-square test of independence.* Biochemia medica: Biochemia medica 23.2 (2013): 143-149.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

[14] Analysis of variance. In Wikipedia, 18 December 2019
https://en.wikipedia.org/wiki/Analysis_of_variance

[15] R LIBRARY CONTRAST CODING SYSTEMS FOR CATEGORICAL VARIABLES. UCLA: Statistical Consulting Group
https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/

[16] Linear regression. In Wikipedia, 20 December 2019
https://en.wikipedia.org/wiki/Linear_regression

[17] Family-wise error rate. In Wikipedia, 11 December 2019
https://en.wikipedia.org/wiki/Family-wise_error_rate

[18] Bonferroni correction. In Wikipedia, 18 December 2019
https://en.wikipedia.org/wiki/Bonferroni_correction

[19] False discovery rate. In Wikipedia, 9 November 2019
https://en.wikipedia.org/wiki/False_discovery_rate

[20] Alzforum - networking for a cure. In Alzpedia, 2020
https://www.alzforum.org/

[21] ABCA7. In Alzpedia, 2020
https://www.alzforum.org/alzpedia/abca7

[22] Complement Receptor 1 (CR1). In Alzpedia, 2020
https://www.alzforum.org/alzpedia/complement-receptor-1-cr1

[23] Karch, Celeste M., and Alison M. Goate. *Alzheimer's disease risk genes and mechanisms of disease pathogenesis.* Biological psychiatry 77.1 (2015): 43-51.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

[24] Liu, Chia-Chen, et al. *Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy.* Nature Reviews Neurology 9.2 (2013): 106.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726719/

[25] Kim, Jungsu, Jacob M. Basak, and David M. Holtzman. *he role of apolipoprotein E in Alzheimer's disease.* Neuron 63.3 (2009): 287-303.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044446/

[26] Random Forest Classifier
https://scikit-learn.org/stable/modules/generated/.sklearn.ensemble.RandomForestClassifier.htmlsklearn.ensemble.RandomForestClassifier

[27] Imbalanced-learn API
https://imbalanced-learn.readthedocs.io/en/stable/api.html
module-imblearn.over$_sampling$

[28] Random Forest Algorithm, An Interactive Discussion
https://www.linkedin.com/pulse/random-forest-algorithm-interactive-discussion-nira

7 Appendix

I. Tables

Table 9. Two-way ANOVA output: combinations of ROI and apoe+picalm genes with significant relationship. Part 1.

Brain area	**adjusted p-value**
left_G_and_S_transv_frontopol	0.0377
left_G_front_sup	0.0498
left_G_occipital_sup	0.0148
left_G_oc.temp_lat.fusifor	0.0432
left_G_oc.temp_med.Lingual	0.0432
left_G_pariet_inf.Angular	0.022
left_G_pariet_inf.Supramar	0.0434
left_G_postcentral	0.0398
left_Lat_Fis.ant.Horizont	0.0432
left_Lat_Fis.ant.Vertical	0.0432
left_S_circular_insula_sup	0.0213
left_S_intrapariet_and_P_trans	0.0057
left_S_oc_middle_and_Lunatus	0.0166
left_S_precentral.inf.part	0.0256
right_G_and_S_frontomargin	0.0005
right_G_and_S_cingul.Mid.Ant	0.0498
right_G_and_S_cingul.Mid.Post	0.0148
right_G_front_inf.Triangul	0.0498
right_G_oc.temp_med.Lingual	0.0286

Table 10. Two-way ANOVA output: combinations of ROI and apoe+picalm genes with significant relationship. Part 2.

Brain area	**adjusted p-value**
right_G_orbital	0.0148
right_G_pariet_inf.Supramar	0.0432
right_G_parietal_sup	0.0432
right_G_subcallosal	0.0498
right_G_temp_sup.Lateral	0.0404
right_G_temp_sup.Plan_tempo	0.0016
right_G_temporal_inf	0.022
right_G_temporal_middle	0.0286
right_Pole_occipital	0.0432
right_S_calcarine	0.0498
right_S_circular_insula_sup	0.0432
right_S_front_middle	0.0498
right_S_front_sup	0.0109
right_S_intrapariet_and_P_trans	0.0377
right_S_oc.temp_lat	0.0432
right_S_orbital.H_Shaped	0.022
right_S_parieto_occipital	0.0432
right_S_precentral.sup.part	0.044
right_S_temporal_inf	0.0432
right_S_temporal_sup	0.0008

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Yevheniia Kryvenko**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Using Machine Learning to Explore Genotype Effects on Cortical Thickness of Human Brain,

   supervised by Raul Vicente Zafra.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Yevheniia Kryvenko
12.03.2020