UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Machine learning and causal inference approaches for systemic multi-disease associations in UK Biobank

Author:
Alejandro GONZÁLEZ ÁLVAREZ

Supervisor:
Dr. Karim LEKADIR
Akshay JAGGI
Dra. Polyxeni GKONTRA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 1, 2020

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Machine learning and causal inference approaches for systemic multi-disease associations in UK Biobank

by Alejandro GONZÁLEZ ÁLVAREZ

Causal inference modeling has captured the attention of many epidemiologists, economists, and machine learning experts in the last few years. The models attempt to discover the causal links between correlated datasets. Rather than just saying "A and B are linked", these models might allow us to say "A causes B" or the other way around. We propose using traditional machine learning techniques as well as new causal inference modeling approaches to mine UK Biobank data to better infer the causal link between multiple heart and brain diseases. We worked with three datasets derived from the UK Biobank: vascular risk factors, that captures how well the heart works, brain MRI indices, structural imaging data that contains the structure of various brain regions, and heart CMR radiomics, that quantify various changes in the heart structure.

Many recent publications have looked at the relationships between single heart and brain diseases. For example, recent work has shown that changes in brain structure correlate with changes in vascular health. Furthermore, past work at the Artificial Intelligence in Medicine Lab at the Universitat de Barcelona has shown that differences in heart CMR radiomics associate with differences in brain imaging. Also, this same lab has shown that changes in heart CMR radiomics correlate with changes in vascular health. However, no work has shown the relative importance and causal links between these three datasets (heart CMR Radiomics, brain MRI indices, and vascular risk factors). Because these connections have been studied independently but not simultaneously, there are potential redundancies in the data. For instance, another group used brain imaging to predict vascular health. However, it's possible that the brain imaging changes associated with vascular health arise from changes in cardiac imaging. Therefore, brain imaging provides no unique information.

Our aim will be to provide the first combined systemic and multi-disease study of all these variables. We used both traditional machine learning techniques and causal inference approaches. For traditional machine learning approaches, we predicted VRFs using brain MRI indices and heart CMR radiomics separately using optimal classifiers. We then combined the two datasets and checked if they performed any better together. If so, we can infer that they provide unique information. After that, we used causal mediation analysis and we assembled graphs of potential relationships between each of the three datasets. We then measured the strength of the connections in these graphs to simultaneously estimate the causal connection between brain diseases, heart diseases, and vascular health.

Acknowledgements

I would like to express my sincere gratitude to anyone who has made possible the development of this master thesis. To my supervisor, Dr. Karim Lekadir, who inspired me to do this master thesis related to artificial intelligence in medicine, a field, and industry with a great impact on society. To my cosupervisors, Akshay Jaggi and Dra. Polyxeni Gkontra, for the help they provided me for the writing and technical parts of this project. Their advice and feedback were highly valuable all throughout the semester, and especially I would like to thank them for their precious time and patience throughout the learning process of this master thesis. To Alejandro Hernandez, my colleague and classmate, for his contribution in part of the code. To master coordinator, Dr. Jordi Vitria, and all the professors in the Master of Fundamental Principles of Data Science at Universitat de Barcelona for their help and guidance throughout the whole academic year. Last but not the least, I would like to thank my family and friends for all their support and encouragement.

Contents

Chapter 1

Introduction

In this chapter, the statement of problems that this thesis aims to solve, motivation, objectives, and structure of the report are described.

1.1 Problem statement

Neurocardiology is a relatively new specialty which was born with the goal of studying and understanding the pathophysiological interplay of the nervous and cardiovascular systems, based on growing evidence that both systems are intertwined. Various pathologies of the nervous system can lead to a wide range of alterations in the function and structure of the cardiovascular system ranging from transient and benign electrographic changes to myocardial injury, cardiomyopathy, and even cardiac death (Tahsili-Fahadan and Geocadin, 2017). However, classical diagnosis and treatment approaches of illnesses have been treated as differentiated and isolated specialties, such as cardiology or neurology. Nonetheless, the human body is a complex network, in which each of the nodes corresponds to the given organ system and the links are representative of the functional interactions between them. These continuous multi-organ interconnections give rise to different physiological states, so that as in any complex network, when any of the nodes or interactions become altered, it may have at some extent an impact over the rest of the systems and thus, over the overall network (Bruns, 2017).

Many observational studies have supported the clinical relevance of multi-organ disease associations, among which we can find many related to the nervous and cardiovascular systems. For example, there are robust data indicating that atherosclerosis and cognitive impairment share common risk factors and pathophysiology (Roger, 2017). Also, heart failure may be induced via sympathetic or parasympathetic activation at certain stressful circumstances (Dokken, 2008), while brain stroke can occur due to atrial fibrillation (Alrabghi et al., 2018). Furthermore, there are as well multi-organ disease associations outside of the heart-brain link. For example, renal dysfunction may be associated with more severe and prevalent cardiovascular complications and mortality, while heart failure can accelerate renal dysfunction (Liu, 2008).

Nonetheless, these single disease associations have been established largely on the basis of epidemiological data, due to insufficient knowledge on the underlying pathophysiologic mechanisms (Pereira et al., 2012). For that reason, to this end of studying the associations between multiple heart and brain diseases, we will take advantage of recent cardiac and brain imaging techniques and their great potential to capture in-vivo small as well as complex changes in both organs and study their inter-relationships.

1.2 Motivation

With new technologies vastly increasing the amount and variety of heart and brain imaging, in conjunction with various cardiovascular disease risk scores, we have plenty of information to study and obtain useful insights from. Combining and gathering this information allows us to discover correlations and associations between multiple heart and brain diseases. The motivation of this thesis is to facilitate the discovery of the causal links between heart structure, brain structure, and vascular health, in addition to compare different traditional and causal approaches.

1.3 Objectives

The main goal of this thesis is to study simultaneously the connections and associations between vascular health, heart structure, and brain structure. Our initial hypothesis is that brain MRI indices and heart CMR radiomics are independent and provide unique information. Therefore, together they will improve performance when we try to predict vascular health using traditional machine learning techniques. However, if this initial hypothesis fails and performance does not improve, it will mean that both datasets do not provide unique information, and as a result, they might be similar and somehow correlated. To study this link between heart and brain structures we will use causal inference techniques such as causal mediation analysis. This thesis is structured around the following specific objectives:

- Reduce the dimensionality of the three datasets provided to work with them in a more efficient manner.
- Extract some relevant and key features from vascular risk factors to better assess and measure vascular health.
- Apply machine learning feature selection algorithms to determine the most important features from heart and brain structures to predict vascular health.
- Run machine learning algorithms to predict vascular health with different combinations of heart and brain structure' variables to test our initial hypothesis.
- Apply propensity score matching techniques to eliminate the possibility that confounding variables are affecting the results, and to be able to have greater confidence that our conclusions are unbiased.
- If our initial hypothesis fails, to study and find the interactions between vascular health, heart structure, and brain structure by using causal inference techniques.
- Compare the performance of traditional machine learning techniques and causal inference approaches in multi-disease association studies.

1.4 Structure of the Report

Chapter 2 describes past related work done and the innovation that this thesis introduces. Next, chapter 3 presents all the theories behind the materials being used. Next, chapter 4 shows all the results from the previously explained techniques. And lastly, chapter 5 presents a summary of the findings, some conclusions, and some limitations and future work that can be done.

Chapter 2

Related work

Plenty of studies have looked at the relationships between single heart and brain diseases, as well as the relationships between heart structure, brain structure, and vascular health individually. In this section, we will review the most important related work and we will introduce our innovative approach.

2.1 Associations between the heart and brain structure's

Predictive modelling of brain structure and function based on cardiovascular magnetic resonance and radiomics

Mireia Masias did her master's thesis with Dr. Karim Lekadir, our supervisor as well, and used heart CMR radiomics to predict the differences in various structures in brain MRI images. To do so, she developed new feature selection algorithms to estimate the best cardiac radiomic predictors of brain measurements and disease (Bruns, 2017).

Cardiac index is associated with brain aging

Although this paper used rudimentary image analysis to generate cardiac index as a measure of vascular health, the author showed that this cardiac index is associated with preclinical brain MRI and neuropsychological markers of ischemia and Alzheimer's disease in the community (Jefferson et al., 2010).

2.2 Associations between brain structure and vascular health

Associations between VRFs and brain MRI indices in UK Biobank

This paper was the editor's choice in 2019 and associations between VRFs and brain structural and diffusion MRI markers were examined in UK Biobank, the same dataset we used. Also, this paper showed that higher levels of VRFs were associated with poorer brain health across grey and white matter macrostructure and microstructure. To study these associations, this paper used many of the techniques we used as well, such as propensity score matching, confirmatory factor analysis, and heart brain connections (Cox et al., 2019a).

Vascular risk burden, brain health, and next steps

Similarly to the previously mentioned article, this article published in 2018 shared the results of an epidemiologic analysis of both cross-sectional and longitudinal data of age-related influences of vascular risk factor burden on brain structure. Both the

cross-sectional and longitudinal analyses found an age-dependent relationship between higher vascular risk burden and lower brain volume. These analyses used an updated version of the Framingham Stroke Risk Profile (FSRP) to assess the vascular risk factor burden, a measure of vascular health (Gorelick and Sorond, 2018). However, as we will explain later, we used two different measures of vascular risk factor burden.

2.3 Associations between heart structure and vascular health

A computational and visualisation tool for investigating associations between cardiac radiomics, risk factors and clinical data

Naphatthara Phloyngamdid also did her master's thesis with Dr. Karim Lekadir, and used heart CMR radiomics features to predict the Framingham Risk Score and other VRFs in UK Biobank patients. To show these associations between cardiac radiomics, risk factors, and clinical data, she developed an intuitive and interactive web-based tool which dynamically displayed the radiomic feature set alongside the additional medical, health, and lifestyle factors feature set based on the contents of radiomics and clinical data files (Phloyngam, 2019).

A radiomics approach to analyze cardiac alterations in hypertension

In this paper, a radiomics approach is described for identifying intermediate imaging phenotypes associated with hypertension, a medical condition that is well established as a risk factor for many major diseases. For example, it can cause alterations in the cardiac structure and function over time that can lead to heart-related morbidity and mortality (Cetin et al., 2020).

2.4 Associations between heart and brain structure and brain cognitive function

Lastly, although is related to this master thesis, the following papers include a new different type of data and variables that we did not study, which corresponds to cognitive and brain function data. For example, the first one studied the associations between cardiometabolic diseases and cognitive abilities, and the results reinforced the notion that preventing or delaying cardiovascular disease or diabetes may delay cognitive decline and possible dementia (Lyall et al., 2016). Moreover, in the second paper of this kind, investigators reported brain structure-intelligence associations on a large sample from the UK Biobank study (Cox et al., 2019b).

2.5 Associations between heart structure, brain structure, and vascular health

As we have seen, there are plenty of interesting papers that have studied the relationships between single heart and brain diseases, as well as the associations between heart structure, brain structure, and vascular health individually. However, no work has followed a multi-disease approach, neither has shown the relative importance and causal links between these three datasets simultaneously as we did in our combined study of all these variables.

Chapter 3

Methodology

3.1 Dataset description: UK Biobank

We worked with three datasets derived from the UK Biobank, a database that comprises longitudinal clinical information and MRI images from a great number of volunteers. This longitudinal long-term study was started in 2006 to provide a detailed database for the improvement in the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses (Biobank, 2020a). This cohort study originally included 500,000 volunteers enrolled at ages comprised between 40 and 69 years old, who are periodically followed up for an expected period of 30 years from its start. In these follow-ups very detailed information about different health-related variables of the volunteer are collected; including genetic information, imaging of heart, brain, abdomen, bones, and carotid artery, as well as biochemistry tests and questionnaires to characterize the cognitive function and daily life (Petersen et al., 2013).

From all this information, in this thesis, we were particularly interested in heart and brain imaging, from which we extracted valuable information regarding their structure, as well as other clinical and daily-life data related to VRFs. Thanks to this database we had access to three complete datasets, which combine added up to a total of 2065 UK Biobank Patients and 1416 variables, measuring vascular health, heart structure, and brain structure:

- VRFs, vascular health data that captures how well the heart works.
- Heart CMR radiomics, that quantify various changes in heart structures.
- Brain MRI indices, structural imaging data that contains the structure of various brain regions.

3.1.1 VRFs

Data in this category are VRFs (vascular risk factors), which aim to measure and quantify vascular health. Each variable in this dataset is a measure of a factor that we know might increase or decrease a patient's risk of developing cardiovascular diseases. These risk factors include age, sex, education status, smoking status, Townsend deprivation (a British measure of poverty), BMI (a measure of weight and obesity), alcohol intake, exercise, diabetes, hypertension, hypercholesterolaemia, cigarette smoking, DBP (diastolic blood pressure), SBP (systolic blood pressure), and WHR (from waist and hip measurements). All these variables were very important in this thesis and we extracted some features from them to better assess vascular health data.

3.1.2 Heart CMR radiomics

Radiomics is a novel image analysis technique, whereby voxel-level information is extracted from digital images and used to derive multiple numerical quantifiers of shape and tissue character - referred to as 'radiomics features' (Raisi-Estabragh et al., 2020), in contrast to the classical approach of treating medical images as pictures intended solely for visual interpretation. In addition, a radiomics study can be structured in five phases: data selection, medical imaging segmentation, feature extraction and quantification, exploratory statistical analysis, and modelling and validation (Lambin et al., 2017). Figure 3.1 presents a workflow showing the study process of radiomics. Furthermore, cardiac magnetic resonance (CMR) is the reference imaging modality for the assessment of cardiac structure and function (Raisi-Estabragh et al., 2020).

In this thesis, we took advantage of radiomics for the acquisition of large amounts of relevant features from CMRs to be able to measure the heart structure. The radiomic feature set was already provided to us. Therefore, for the purpose of this thesis, no feature extraction was required. These features typically fail to be appreciated by the naked eye and are commonly grouped into three main categories: first order, second order and higher-order features (Gillies, Kinahan, and Hricak, 2016). These features can be further classified into more subcategories and an overview of them can be seen in Table 3.1.

- First-order statistics describe the distribution of values of individual voxels without concern for spatial relationships. These are generally histogram-based methods and reduce a region of interest to single values for mean, median, maximum, minimum, and uniformity or randomness (entropy) of the intensities on the image, as well as the skewness (asymmetry) and kurtosis (flatness) of the histogram of values.
- Second-order statistical descriptors generally are described as "texture" features, and they describe statistical interrelationships between voxels with similar (or dissimilar) contrast values.
- Higher-order statistical methods impose filter grids on the image to extract repetitive or non-repetitive patterns.

3.1.3 Brain MRI indices

Our brain is composed of two relevant tissue types: gray and white matter. Gray matter is where most of the computation in brains happen. On the other hand, scientists talk about white matter tracts, which are bundles of axons that connect the different brain regions. In the plot (Figure 3.2), part A is all the gray matter regions and part B is all the white matter tracts. Therefore, there are two main parts of the brain that the different MRIs modalities aim to quantify: the structure and size of gray matter regions, and the structure and size of white matter pathways. There are multiple kinds of structural MRI data in the UK Biobank, and for this thesis, we were provided with the following structural MRI data (Biobank, 2020b):

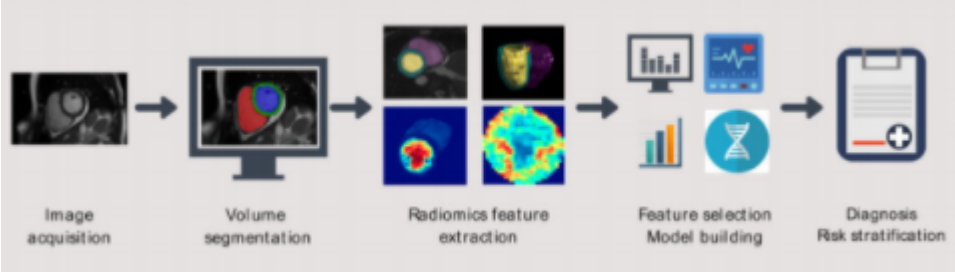

FIGURE 3.1: (1) Image acquisition. (2) Volume segmentation: The areas to be analysed are contoured. (3) Radiomics feature extraction: Radiomics features are extracted from the segmented region of interest. (4) Feature selection: Features that are most robust and informative are selected from the extracted features using different methods. (5) Model building: The selected radiomics features are used as predictor variables to build statistical models for disease discrimination or outcome prediction. (6) Diagnosis, risk stratification: Models undergo internal and external validation and may ultimately be incorporated into clinical care for improved diagnostic accuracy and/or outcome prediction (Raisi-Estabragh et al., 2020).

TABLE 3.1: Radiomics features overview (Martin-Isla et al., 2020)

Type	**Description**	**Examples**
Shape features	Describe geometric characteristics of the cardiac structures	Volume, surface area, sphericity, diameters, axis, surface to volume ratio, flatness
Intensity (First order)	Statistics on the intensity distributions within the ROI	Mean intensity, range, skewness (asymmetry) and entropy
Texture GLCM (Second order)	Quantifies the spatial relationship of the pixels in the ROI	Contrast, correlation
Texture GLSZM (Higher order)	Quantifies the number of connected voxels that share the same intensity level	Gray level non-uniformity, zone entropy
Texture GLRLM (Higher order)	Quantifies the gray level runs in the ROI	Run entropy, long run emphasis and short run emphasis
Texture NGTDM (Higher order)	Quantifies the difference between a gray value and the average gray value of its neighbors within a predefined distance	Busyness, strength
Texture GLDM (Higher order)	Quantifies the gray level dependencies in the ROI	Dependence nonuniformity, dependence entropy and dependence variance
Fractal dimension	Determines the ratio of change in detail to the change in scale	

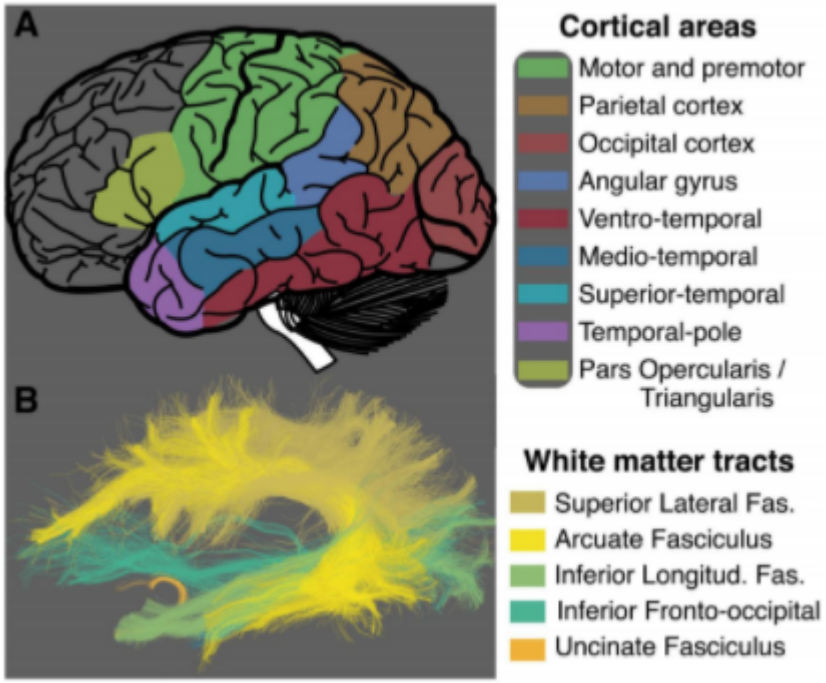

FIGURE 3.2: Structure of the brain.

- T1-weighted structural MRI, high-resolution measurements of the volumes of various regions.
- T2-FLAIR MRI, a good measure of white matter pathologies, changes in the white matter intensity.
- T2*, a measure that is sensitive to microstructures containing potentially useful molecules like myelin and iron.
- FA, fractional anisotropy, a measure of white matter integrity.
- MD, mean diffusivity, the average strength of water diffusion in the brain.
- MO, diffusion tensor mode, how uniform white matter is in the brain, ordered vs disordered.
- L1,2,3, the strength of the diffusion in each cartesian direction x,y,z.
- ICVF, intra-cellular volume fraction.
- ISOVF, isotropic volume fraction.
- OD, orientation dispersion index.

3.2 Data pre-processing

The datasets we were given, described above, were previously cleaned and delivered to us afterwards. Therefore, not a lot of data pre-processing was required, besides the following actions:

- Merging and combining the datasets provided according to patients ids.
- Selecting data instance collection. The same data was collected in three instances. We selected the instance with the least number of missing values and drop the other two.
- Remove missing values. Due to our large sample size, removing entire rows with missing values was considered to be the best approach to have a complete final dataset.
- Substituting "Do not know" responses. Cigarette smoking variables presented plenty of these types of categorical values. Therefore, replacing them with 0 was considered to be the best approach.
- Encoding categorical variables. Similarly, as the previous example, some other categorical variables were encoded using integer encoding, that consist in referencing each possible categorical value with an integer, such as "Less than one a day" substituted by 1.

3.3 Feature Extraction

Several new features were extracted from our initial VRFs variables:

- Smoking pack years. Pack years was calculated as the number of cigarettes per day smoked by each patient, then divided by 20, and lastly multiplied by the number of years patients reported having smoked for (Cox et al., 2019a).
- WHR (Waist Hip Ratio). Waist and hip measurements were conducted to provide WHR.
- BMI (Body Mass Index). BMI was calculated as

$$\frac{weight(kg)}{height(m)^2} \tag{3.1}$$

3.4 Data normalization

Due to the diverse nature of different information sources in cardiac medicine, a normalization step is often required before model crafting. In general, learning algorithms benefit from standardization of the data set, and for example, some algorithms will improve cardiovascular predictions if all numerical features are zero centered and have a variance of the same magnitude order (Martin-Isla et al., 2020). For this reason, some of our VRFs were normalized using Z-Score normalization due to the nature and distribution of the variables. These variables were all used, in addition to some other binary variables, to derive our aggregate measure of vascular risk and to compute our latent factor of general vascular risk (gVRF). These variables

were SBP, DBP, Pack years, BMI, and WHR, and the formula for calculating Z-Score normalization is the following, where x is the raw score, μ is the population mean, and σ is the population standard deviation:

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

Also, before running the ML algorithms that we will explain later, we also used "MinMaxScaler" from the Scikit-learn package in python to normalize all of our heart CRM radiomics and brain MRI indices. This normalization technique transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one (Pedregosa et al., 2011). The formula for calculating "MinMaxScaler" normalization is the following:

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{3.3}$$

3.5 Dimensionality reduction

One of the main objectives of this thesis was to reduce the dimensionality of the three datasets provided to work with them more efficiently. Since we were working with 2065 UK Biobank Patients and 1416 variables, this step played an important role in the development of the rest of the thesis. For this reason, first of all, we derived an aggregate measure of vascular risk to quantify the overall load of VRFs for each patient. Secondly, we also derived a latent factor of general vascular risk (gVRF) to capture the tendency for VRFs to co-occur. And lastly, we also derived several latent factors from heart CMR radiomics and brain MRI indices using both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA).

3.5.1 Aggregate measure of vascular risk

An aggregate measure of vascular risk for each individual was derived to quantify the overall load of VRFs for each patient, counting instances of a self-reported diagnosis of hypertension, diabetes, or hypercholesterolaemia, having ever smoked, having a BMI >25, and having a high WHR (>0.85 for females and >0.90 for males), following prior work in an older cohort (Cox et al., 2019a).

3.5.2 gVRF

We also derived a latent factor of general vascular risk (gVRF) following prior work in an older cohort (Wardlaw et al., 2014), using confirmatory factor analysis in structural equation modelling. This latent measure captures, and depends upon, the tendency for VRFs to co-occur. Using the Factor Analyzer package for python, gVRF was derived from smoking pack years, DBP, SBP, BMI, WHR, self-reported hypertension, diabetes, and hypercholesterolaemia.

3.5.3 SelectKbest feature selection

In addition to the previously mentioned derived latent factors and measures, to reduce the dimensionality of our datasets, before running the ML algorithms that we will explain later, we also selected the best five features according to the highest scores computed by the ANOVA F-value from the "SelectKbest" and "feature-selection" classes in Scikit-learn. These methods based on the F-test estimate the degree of linear dependency between two random variables (Pedregosa et al., 2011). To do so, a statistical F-test uses an F Statistic to compare two variances, $s1$ and $s2$, by following the next equation:

$$F = \frac{s_1^2}{s_2^2} \tag{3.4}$$

3.6 Factor Analysis

Factor analysis is a multivariate statistical approach used to search influential latent variables from a set of observed variables. This dimensionality reduction technique helped us in the process of data interpretation by reducing a large number of variables from our datasets into a smaller set of variables, also known as factors. Factor analysis is widely utilized in market research, psychology, operations research, and more recently in health-related professions, where it has become much more common during the past two decades. This increase is illustrated in recent surveys of health science electronic databases, where articles reporting factor analysis increased by 16 percent (Williams, Onsman, and Brown, 2010).

There are two main types of factor analysis: Exploratory Factor Analysis (EFA), and Confirmatory Factor Analysis (CFA). The basic assumption behind EFA is that any observed variable is directly associated with any factor. On the other hand, the basic assumption behind CFA is that each factor is associated with a particular set of observed variables. Lastly, the main advantage of factor analysis is that the extracted factors are interpretable thanks to the factor' loadings. Loadings determine the strength of the relationships, and factors can be identified by the largest loadings (Yong and Pearce, 2013). We conducted both Exploratory and Confirmatory Factor Analysis (EFA and CFA) using the Factor Analyzer package for python to derive several latent factors from heart structure and brain structure datasets, in addition to the previously mentioned gVRF. To measure the suitability of FA, three different tests were run: KMO Test, Bartlett's Test, and Scree test.

KMO Test

Kaiser-Meyer-Olkin (KMO) Test measures the suitability of data for factor analysis since it determines the adequacy for each observed variable and the complete model. It estimates the proportion of variance among all the observed variables and a lower proportion is more suitable for factor analysis. KMO values range between 0 and 1, with 0.50 and above considered suitable for factor analysis (Williams, Onsman, and Brown, 2010). The formula for the KMO test is the following:

$$MO_j = \frac{\Sigma_{i \neq j} \, r_{ij}^2}{\Sigma_{i \neq j} \, r_{ij}^2 + \Sigma_{i \neq j} \, u} \tag{3.5}$$

In this formula, $R = [r_{ij}]$ is the correlation matrix and $U = [u_{ij}]$ is the partial covariance matrix (Cerny and Kaiser, 1977).

Bartlett's Test

Bartlett's test of sphericity checks whether or not the observed variables intercorrelate at all using the observed correlation matrix against the identity matrix. If the test is found to be statistically insignificant, FA should not be employed. On the contrary, if the p-value is p<.05, the test is considered statistically significant, indicating that the observed correlation matrix is not an identity matrix, and the data tested is suitable for FA (Williams, Onsman, and Brown, 2010).

Eigenvalues, Kaiser's criterion, and Scree plot

Eigenvalues, also known as characteristic roots, represent the variance explained by each factor from the total variance. Kaiser criterion is an analytical approach, which is based on that the more significant proportion of variance explained by the factor will be selected. The eigenvalue approach is a good criterion for determining the number of factors. Generally, an eigenvalue greater than 1 will be considered as a selection criterion. Furthermore, the graphical approach is based on the visual representation of factors' eigenvalues, also called a scree plot. This scree plot helps us to determine the number of factors where the curve makes an elbow, and this decision involves two steps: first, drawing a straight line through the smaller eigenvalues where a departure from this line occurs, highlighting this way where the break occurs, and second, the point above this break indicates the number of factors to be retained (Williams, Onsman, and Brown, 2010).

3.7 Machine Learning

Machine learning (ML) approaches to image-based diagnosis rely on data-driven algorithms that learn from past clinical examples through the identification of hidden and complex imaging patterns. Existing work already demonstrates the incremental value of image-based cardiovascular diagnosis with ML for several important conditions such as coronary artery disease (CAD) and heart failure (HF) (Martin-Isla et al., 2020). The overall pipeline to build ML tools for image-based cardiac diagnosis is described in Figure 3.3. The advantage of machine learning with respect to traditional statistical methods lies in its ability to use features that have long been excluded in the decision making due to ignorance of their relationship to the condition of interest. Combined with machine learning, biomedical images offer a great source of clinical information. Much valuable information is contained in images, which include not only those standard imaging indices which physicians rely upon, but also highly valuable patterns invisibly to the untrained human eye.

The radiomics approach in combination with ML is proposed in this thesis specifically to enable deeper characterization of cardiac imaging predictors that can be used to the study of heart-brain links (Bruns, 2017). To apply ML in this thesis, we predicted vascular health using brain MRI indices and heart CMR radiomics separately using optimal classifiers. We then combined the two datasets and saw if they

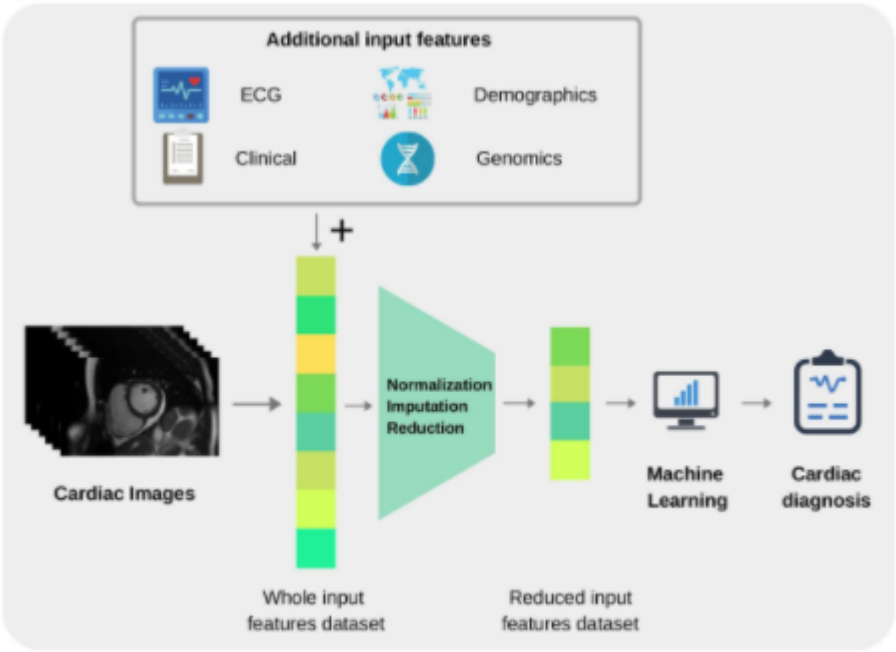

FIGURE 3.3: ML pipeline for image-based cardiac diagnosis. In short, it requires (1) input imaging datasets from which suitable imaging predictors can be extracted, (2) accurate output diagnosis labels, and (3) a suitable ML technique that is typically chosen and optimized depending on the application to predict the cardiac diagnosis (output) based on the imaging predictors (input) (Martin-Isla et al., 2020).

performed any better together. If so, we can infer that they provide unique information. Specifically, we predicted gVRF and our aggregate measure of vascular risk.

3.7.1 Predicting gVRF

Our latent factor of general vascular risk (gVRF) is a continuous variable. Therefore, we considered that the optimal algorithm to predict it is the Ordinary Least Squares Linear Regression from the "linear-model" class in Scikit-learn.

Linear regression

This algorithm is a linear model with coefficients $w = (w1, \ldots, wp)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation (Pedregosa et al., 2011).

3.7.2 Predicting aggregate measure of vascular risk

Our aggregate measure of vascular risk is a discrete categorical variable. Therefore, we considered that the optimal algorithm to predict it is a supervised learning model for the classification of the aggregate measure, which values range from 0 to 5. Also, due to the imbalanced nature of our target variable, we considered that the best algorithm to choose is Random Forest (RF) from the "ensemble" class in Scikit-learn (Pedregosa et al., 2011). Furthermore, to reduce the imbalanced data effect in our target variable we applied several different techniques such as Random Oversampling, Random Undersampling, and SMOTE-Tomek.

Random Forest

This popular technique consists of a combination of decision trees (DTs) trained on different random samples of the training set. Each DT is a set of rules based on the input features values optimized for accurately classifying all elements of the training set. DTs are nonlinear models and tend to have high variance. If the DT is grown very deep it can pick up irregularities in the training dataset and consequently problems with overfitting may be encountered. This problem is counteracted in a RF through training on different samples of the training dataset. In this way the variance is reduced as the number of DT used, lowering therefore the generalization error and becoming a powerful technique. The final prediction is obtained by selecting the mode for classification problems of all predictions (Martin-Isla et al., 2020).

Random Undersampling

Random undersampling involves randomly selecting examples from the majority class to delete from the training dataset. This has the effect of reducing the number of examples in the majority class in the transformed version of the training dataset. This process can be repeated until the desired class distribution is achieved, such as an equal number of examples for each class. The random undersampling technique was implemented using the "resample" class in Scikit-learn (Pedregosa et al., 2011).

Random Oversampling

Random oversampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new “more balanced” training dataset multiple times. The random oversampling technique was implemented as well using the "resample" class in Scikit-learn (Pedregosa et al., 2011).

SMOTE-TOMEK

SMOTE (Synthetic Minority Oversampling Technique) consists of synthesizing elements for the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors. On the other hand, Tomek links are pairs of very close instances, but of opposite classes. Removing the instances of the majority class of each pair increases the space between the two classes, facilitating the classification process.

So in summary, SMOTE-TOMEK is a combination of oversampling and undersampling techniques, where SMOTE generates noisy samples by interpolating new points between marginal outliers and inliers, and Tomek’s cleans the space resulting from oversampling. This SMOTE-TOMEK technique was implemented using the imblearn library for python (Lemaître, Nogueira, and Aridas, 2017).

TABLE 3.2: Linear regressions' performance metrics

Metric	Description
R^2	Coefficient of determination, regression score function
MAPE	Mean absolute percentage error
MAE	Mean of the absolute value of the errors
MSE	Mean of the squared errors
RMSE	Square root of the mean of the squared errors

TABLE 3.3: Random Forest' performance metrics

Metric	Description
Accuracy	Measures the percentage of the algorithm classifying the input data correctly
Confusion Matrix	Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class
ROC Curve	Performance plot representation created by plotting the true positive rate against the false positive rate at various threshold settings
AUC	When a trained model is asked to make a prediction, a probability can be computed and used to generate a ROC analysis

3.7.3 Validation

For the validity of the algorithms, we split our input data into two different subgroups, called training set and testing set, respectively. Once the ML models were trained and tested, we used different metrics to evaluate their performance. In the case of the Ordinary Least Squares Linear Regression predicting gVRF, the following metrics on Table 3.2 were used. On the other hand, the metrics on Table 3.3 were used for the Random Forests predicting our aggregate measure of vascular risk.

3.8 Propensity Score Matching

Propensity score matching analysis was applied to eliminate the possibility that confounding variables are affecting the results, and to be able to have greater confidence that our conclusions are unbiased. This method identifies similar patients and studies the difference in those confounding variables between only those patients. Usually, propensity score matching is used when a group of subjects receive a treatment and we'd like to compare their outcomes with the outcomes of a control group. In this thesis, patients with a total of four or five VRFs were each matched on sex and age to a single participant who had no VRFs. The essential idea behind this theory is to create a balance among the groups of interest on important covariates. With this technique, a single variable is computed, a propensity score, that captures how differences in these variables contribute to a subject's statistical probability of being in one group or another. This term, propensity score, really refers to a kind of index or composite variable that summarizes important group differences. Subjects with similar propensity scores resemble each other with respect to these characteristics and those with very different propensity scores are dissimilar. Propensity scores are typically computed using logistic regression, with group (treatment) status regressed on observed baseline characteristics such as age, gender, and behaviors of relevance to the research. Then, these propensity scores are the predicted probabilities of being in one group or another that have been derived from the model. In addition, propensity scores can be used to create matched samples with both one-to-one matching or one-to-many matching.

Therefore, in summary, propensity score matching is done in two steps: first, estimating the propensity score, the conditional probability of receiving treatment given the covariates, and secondly matching pairs with the same (or similar) propensity scores (Ottoboni, 2020). To do so, we did our matching with replacement and allowed one control to be matched to one or more cases because matching with replacement minimizes the propensity score distance between the matched comparison units and the treatment unit, which is beneficial in terms of bias reduction (Dehejia and Wahba, 1998). To apply these propensity score matching techniques we used two python packages: pscore match and pymatch.

3.9 Causal Inference

In addition to the previously mentioned ML models and propensity score matching analysis, for this thesis we also proposed using new causal inference modeling techniques to mine the UK Biobank data to better infer the causal link between heart and brain diseases. For that purpose, we used causal mediation analysis, we assembled several graphs of potential relationships between each of the three datasets, and measured the strength of the connections in these graphs to simultaneously estimate the causal connection between brain structures, heart structures, and vascular health.

As we have already discussed, many recent publications have proved that changes in brain structure correlate with changes in vascular health, differences in heart CMR radiomics are associated with differences in brain imaging, and changes in heart CMR radiomics correlate with changes in vascular health. However, because these

TABLE 3.4: Mediation Analysis performance metrics

Metric	Description
ACME (average causal mediation effect)	Total effect minus the direct effect (TE - ADE)
ADE (average direct effect)	A direct effect of X on Y after taking into account a mediation indirect effect of M ($X + M \rightarrow Y$)
TE (total effect)	Indirect + direct effect. A total effect of X on Y (without M) ($X \rightarrow Y$)

connections have been studied independently but not simultaneously, there are potential redundancies in the data. For this reason, causal mediation analysis plays an essential role by helping to identify intermediate variables (or mediators) that lie in the causal pathway between the treatment and the outcome (Imai, Keele, and Tingley, 2010). Furthermore, the goal of researchers is not only estimating the causal effects of a treatment but also understanding the process in which the treatment causally affects the outcome. Causal mediation analysis is frequently used to assess potential causal mechanisms (Tingley et al., 2014).

To apply these causal mediation analyses we used the "Mediation" class from the "Statsmodels" library, the Python version for the "mediation R package". This package implements a comprehensive suite of statistical tools for conducting such an analysis and is organized into two distinct approaches. For this thesis, we used the model-based approach, in which researchers can estimate causal mediation effects and conduct sensitivity analysis under the standard research design (Tingley et al., 2014). In Figure 3.4 we can see a generic graphical representation of a mediation analysis, in which three sets of regressions can be seen:

$$X \rightarrow Y \tag{3.6}$$

$$X \rightarrow M \tag{3.7}$$

$$X + M \rightarrow Y \tag{3.8}$$

First of all, we want X to affect Y. If there is no relationship between X and Y, there is nothing to mediate. Secondly, we also want X to affect M. If X and M have no relationship, M is just a third variable that may or may not be associated with Y. Lastly, we want M to affect Y, but X to no longer affect Y, or X to still affect Y but in a smaller magnitude. If a mediation effect exists, the effect of X on Y will disappear (or at least weaken) when M is included in the regression. The effect of X on Y goes through M. If the effect of X on Y completely disappears, M fully mediates between X and Y (full mediation). If the effect of X on Y still exists, but in a smaller magnitude, M partially mediates between X and Y (partial mediation). In Table 3.4 the metrics being used to assess the performance of these models are included.

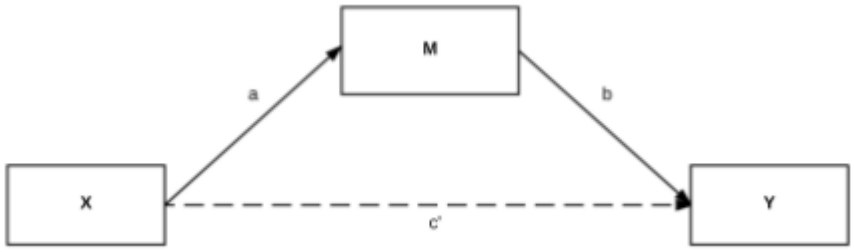

FIGURE 3.4: Generical graphical representation of a mediation analysis. a and b reflect the indirect path of the effect of X on the outcome (Y) through the mediator (M), while c' is the direct effect of X on the outcome after the indirect path has been removed. The total effect of X is the combined indirect and direct effects (Clark, 2019).

Chapter 4

Results

4.1 Data normalization

In Figure 4.1 we can see the distribution of VRFs before normalization. On the other hand, in Figure 4.2 we can see the distribution of the VRFs that were normalized using Z-Score normalization. These variables are SBP, DBP, Pack years, BMI, and WHR. As we can see, all these variables are centered at 0 and have a standard deviation of 1, handling this way all the outliers that were present in the data. Also, it is noticeable that blood pressure-related variables (SBP and DBP) have a normal distribution.

4.2 Dimensionality reduction

4.2.1 Aggregate measure of vascular risk

In Figure 4.3 we can see the distribution of our derived aggregate measure of vascular risk. This measure quantifies the overall load of VRFs for each patient, and as we can see the distribution is very imbalanced, since for example there are only 17 patients with an overall load of VRFs of 5, while there are 606 patients with an overall load of VRFs of 1. This imbalanced target variable will present us with some complications when we try to apply some ML algorithms to predict it.

4.2.2 Latent factor of general vascular risk (gVRF)

In Figure 4.4 we can see the distribution of our derived gVRF, which is as well centered at 0 and has a normal distribution, similar to the variables to which this latent factor was derived with. Also, the KMO test result and the Bartlett's test of sphericity, which can be seen in Table 4.2, are considered adequate and statistically significant. Therefore, securing the suitability of the data being used to extract this latent factor for FA. Furthermore, In Figure 4.5 we can see a correlation plot of all the variables being used to extract this gVRF, in which a strong correlation between SBP and DBP is noticeable, as well as for BMI and WHR. This result makes sense since these pairs of variables measure similar VRFs. Lastly, in Table 4.1 we can see the loadings of the variables being used. The factor loading is a matrix that shows the relationship of each variable to the underlying factor. It also shows the correlation coefficient for the observed variables and the factor, and the variance explained by the observed variables. Loadings are inconsistent (range 0.1557–0.7232), with the factor more strongly loaded towards DBP and SBP.

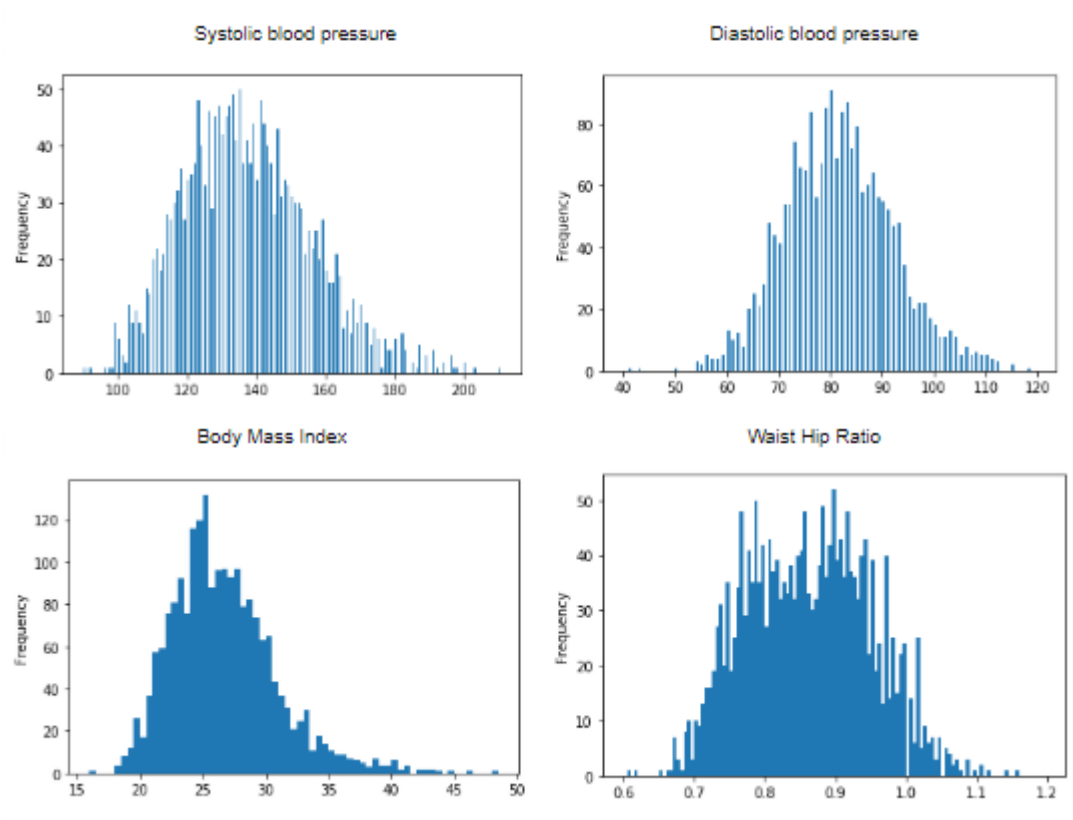

FIGURE 4.1: VRFs before Z-Score Normalization

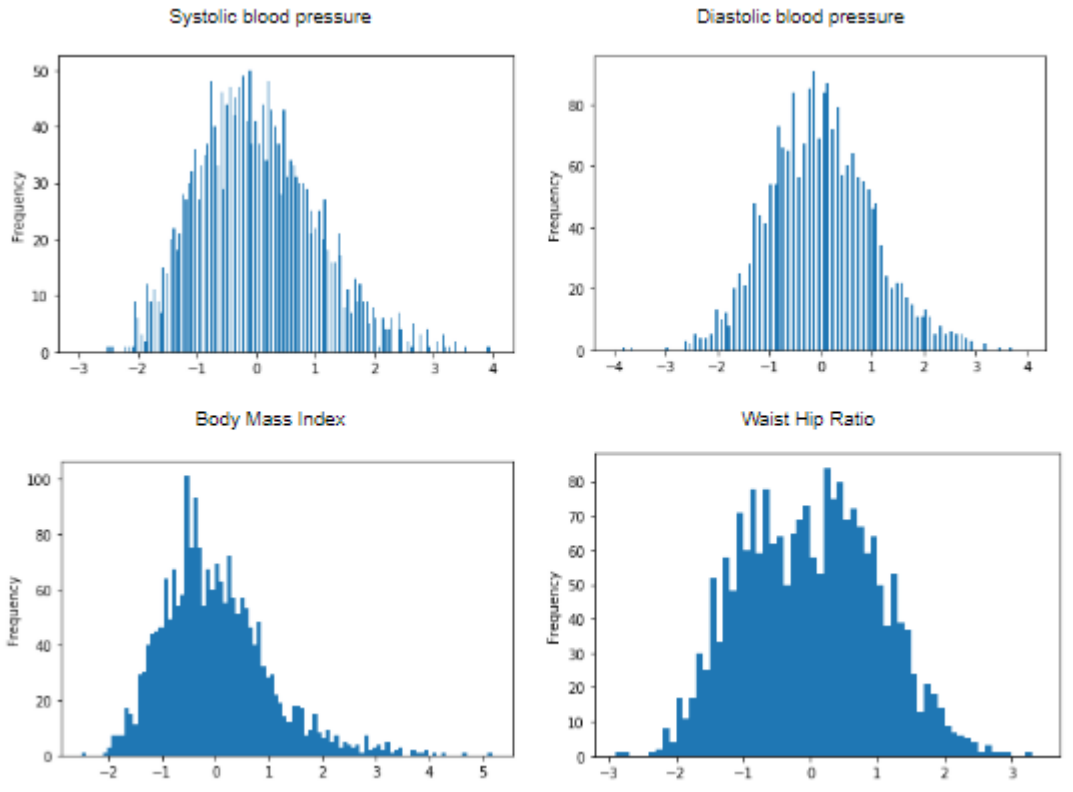

FIGURE 4.2: VRFs after Z-Score Normalization

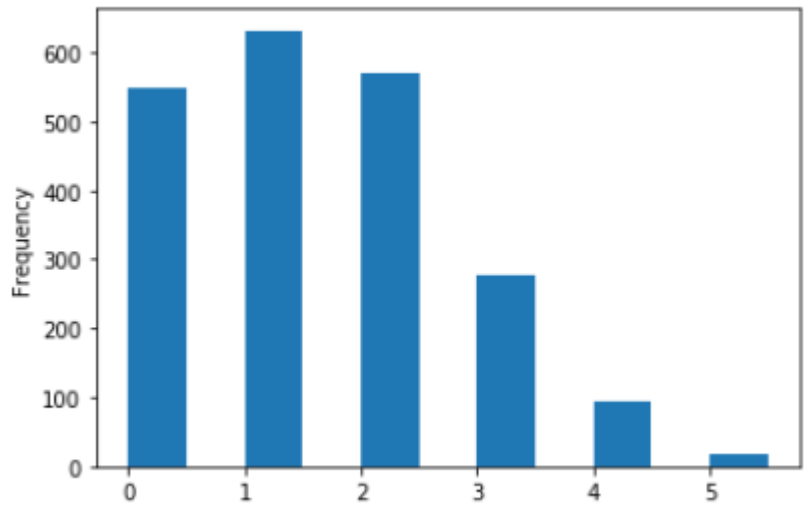

FIGURE 4.3: Aggregate measure of vascular risk

TABLE 4.1: gVRF variables' loadings

Variables	Loadings
DBP	0.723294
SBP	0.686705
WHP	0.490945
BMI	0.457422
Hypertension	0.380753
Hypercholesterolaemia	0.283791
Pack Years	0.186169
Diabetes	0.155718

4.3 Factor Analysis

We conducted both EFA and CFA using the Factor Analyzer package for python to derive several latent factors from heart structure and brain structure datasets, in addition to the previously mentioned gVRF. The KMO test result which can be seen in Table 4.2, is considered adequate. However, the output from the Bartlett's test of sphericity was not statistically significant. Furthermore, we also conducted the Kaiser criterion and the eigenvalue approach to determine the ideal number of factors for each dataset. In Figure 4.6 we can see the derived scree plots, where the ideal number of factors for heart CMR radiomics is set around 50, and for brain MRI indices around 100. Lastly, in Table 4.3 and Table 4.4 we can see the top ten loadings of the variables being used for each factor. For heart CMR radiomics, the factor is more strongly loaded towards Texture GLCM (second-order), Texture GLRLM (higher-order), and Texture GLDM (higher-order) features. On the other hand, in case of brain MRI indices, factor is more strongly loaded towards L3, MD, and L2 features.

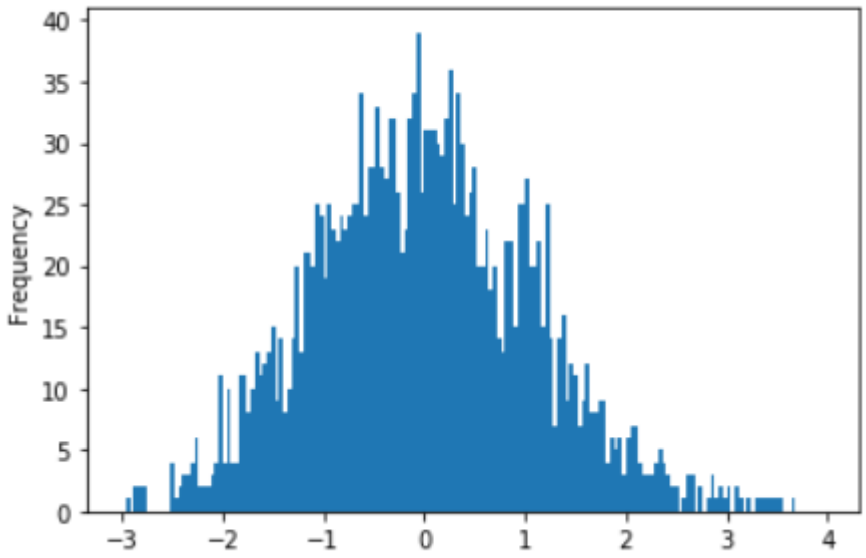

FIGURE 4.4: Latent factor of general vascular risk (gVRF)

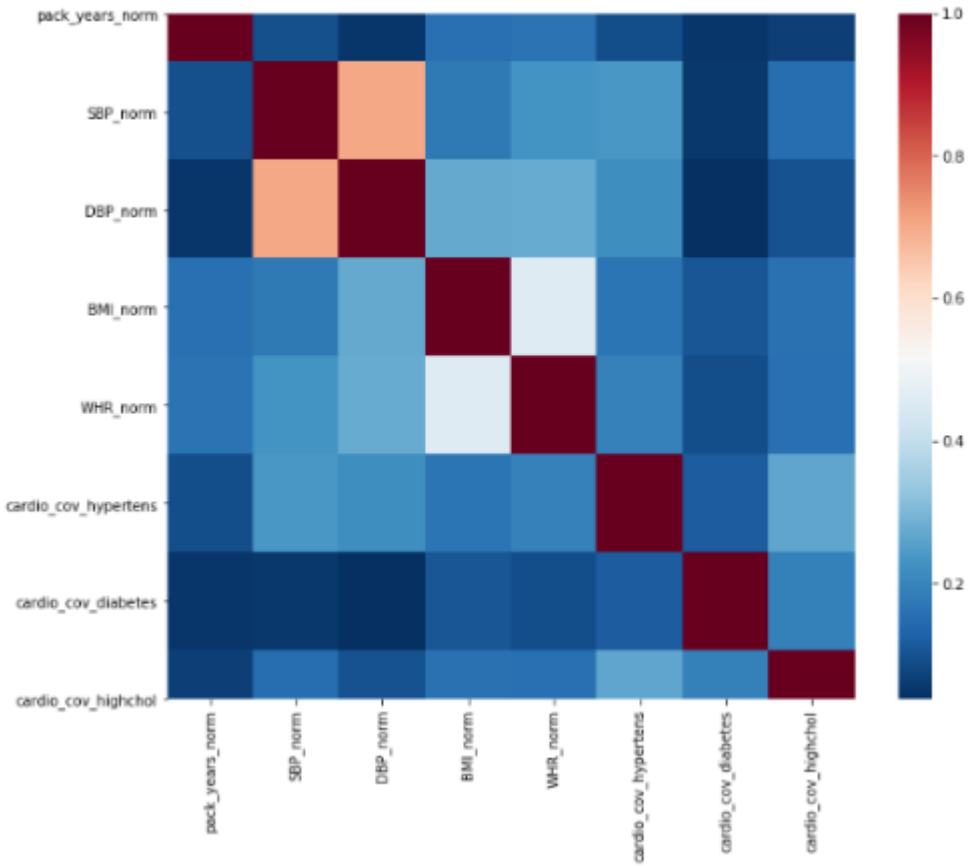

FIGURE 4.5: Correlation plot of gVRF

TABLE 4.2: VRFs, Brain and Heart latent factors

Features	Variables	KMO Score	Bartlet's test	Scree test
gVRF	8	0.6418	(2719.71, 0)	3
Brain MRI Indices	744	0.9526	(inf, nan)	100
Heart CMR Radiomics	639	0.9781	(inf, nan)	50

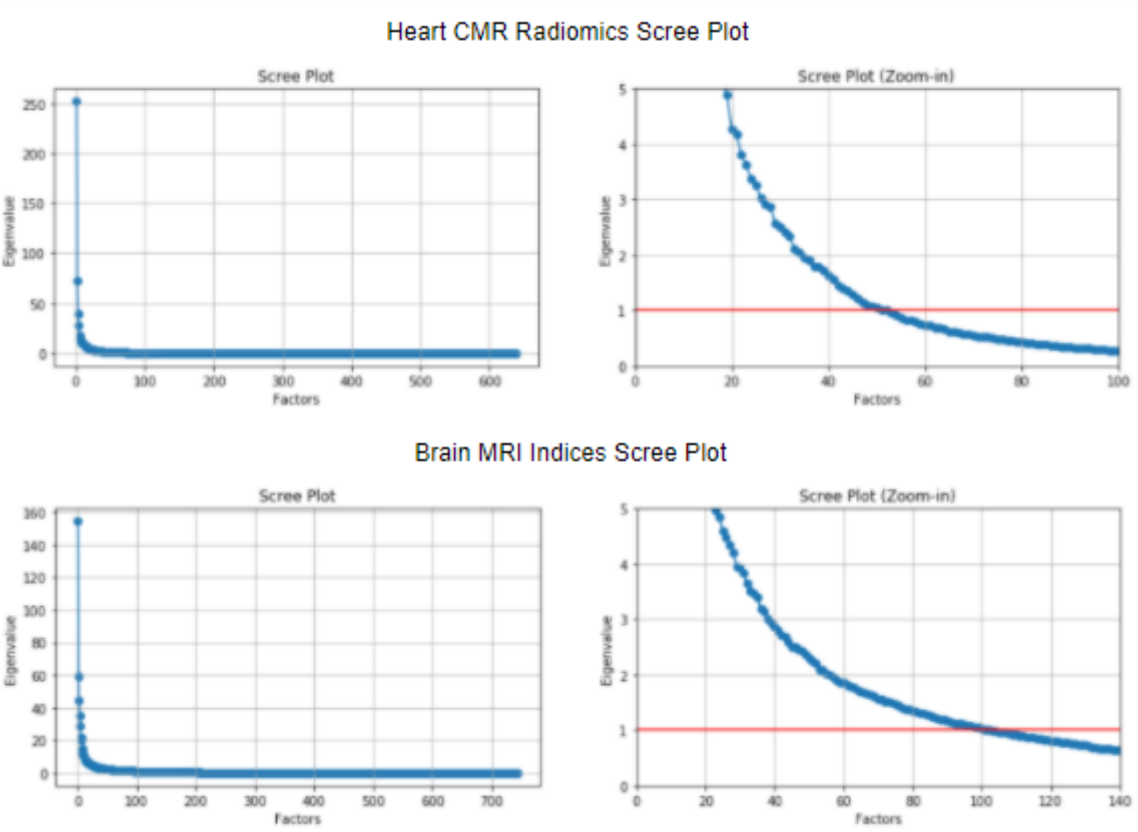

FIGURE 4.6: Scree plots for heart and brain datasets

TABLE 4.3: Top ten Heart CMR Radiomics variables' loadings

Variables	Loadings
Heart Inverse Difference glcm RV ES texture	0.891137
Heart Inverse Difference Moment glcm RV ES texture	0.889092
Heart Inverse Difference glcm RV ED texture	0.880894
Heart Inverse Difference Moment glcm RV ED texture	0.880765
Heart Gray Level Non Uniformity Normalized glrlm RV ES texture	0.871713
Heart Large Dependence Low Gray Level Emphasis gldm RV ED texture	0.867649
Heart Gray Level Non Uniformity Normalized glrlm LV ES texture	0.866858
Heart Gray Level Non Uniformity Normalized glrlm RV ED texture	0.858650
Heart Inverse Difference glcm LV ES texture	0.858577
Heart Large Dependence Low Gray Level Emphasis gldm LV ED texture	0.856550

TABLE 4.4: Top ten Brain MRI Indices variables' loadings

Variables	Loadings
Brain mean l3 in anterior corona radiata on fa skeleton left	0.875992
Brain mean l3 in anterior corona radiata on fa skeleton right	0.870006
Brain mean md in anterior corona radiata on fa skeleton left	0.865722
Brain weighted mean l3 in tract inferior fronto occipital fasciculus right	0.859523
Brain mean md in superior longitudinal fasciculus on fa skeleton left	0.858544
Brain mean l2 in anterior corona radiata on fa skeleton left	0.857454
Brain mean md in anterior corona radiata on fa skeleton right	0.856520
Brain mean l3 in superior corona radiata on fa skeleton left	0.853679
Brain mean md in superior longitudinal fasciculus on fa skeleton right	0.852527
Brain mean l3 in superior longitudinal fasciculus on fa skeleton right	0.851144

4.4 Machine Learning

4.4.1 Predicting gVRF

We obtained the following results when predicting our gVRF with ordinary least squares linear regressions:

- In Table 4.5 we can see all the error metrics for the linear regressions when using FA as the dimensionality reduction technique (we selected 5 latent factors for each group of variables).
- In Table 4.6 we can see all the error metrics for the linear regressions when using SelectKbest as the dimensionality reduction technique (we selected the 5 SelectKbest features for each group of variables).
- In Figure 4.7 we can see the plots for R^2 and MAPE for both dimensionality reduction techniques.
- In Figure 4.8 we can see the plots for MAE, MSE, and RMSE with FA as the dimensionality reduction technique.
- In Figure 4.9 we can see the plots for MAE, MSE, and RMSE with SelectKbest as the dimensionality reduction technique.

R^2 is a goodness-of-fit measure for linear regression models. With our results with can see how the best R^2 using FA is with the heart radiomics dataset. On the contrary, using SelectKbest, the best R^2 is with the combination of cardio CMR and brain MRI indices datasets, which is a little bit higher than brain MRI indices by itself. In the case of MAPE, a commonly used loss function for regression problems, the best results, those where this error metric is the lowest, are obtained with brain MRI indices using both dimensionality reduction techniques. Lastly, in the case of MAE, MSE, and RMSE, the best results, those where these error metrics are minimum, are obtained with heart radiomics in the case of FA, and with brain MRI indices and its combination with cardio CMR in the case of SelectKbest features selection.

However, overall, the combination of these datasets does not seem to improve at all or improve very little our performance metrics, which may lead us to reject our initial hypothesis. Therefore, we can infer that these datasets do not provide unique

TABLE 4.5: Results predicting gVRF with FA

Variables	R^2	MAPE	MAE	MSE	RMSE
Heart Radiomics	0.2183	3.1363	0.7892	0.9565	0.9780
Cardio CMR	0.1483	2.3338	0.8211	1.0422	1.0208
Brain MRI Indices	0.1297	1.9983	0.8236	1.0650	1.0319
Heart Radiomics and Brain MRI Indices	0.1492	2.3339	0.8200	1.0410	1.0203
Cardio CMR and Brain MRI Indices	0.1297	1.9983	0.8236	1.0650	1.0319

TABLE 4.6: Results predicting gVRF with SelectKbest

Variables	R^2	MAPE	MAE	MSE	RMSE
Heart Radiomics	0.2715	3.0125	0.7656	0.8914	0.9441
Cardio CMR	0.1486	2.2747	0.8211	1.0418	1.0207
Brain MRI Indices	0.2824	1.9459	0.7294	0.8780	0.9370
Heart Radiomics and Brain MRI Indices	0.2715	3.0125	0.7656	0.8914	0.9441
Cardio CMR and Brain MRI Indices	0.3280	1.9703	0.7167	0.8222	0.9068

information, and somehow they are related. In Figure 4.10 we can see a correlation plot between the top five SelectKbest selected features from the heart and brain datasets, in which a strong correlation can be seen between heart and brain features. We were then interested in how the prediction of our continuous measure of vascular risk compares to prediction of our discrete measure.

4.4.2 Predicting aggregate measure of vascular risk

As we already mentioned before, our aggregate measure of vascular risk is a very imbalanced target variable as we can see in Figure 4.3. Therefore, even though we tried to predict this class treating the problem as a multi-class classification problem, results were not very satisfactory, and we achieved very low accuracies and bad predictions. The main reason is due to the imbalanced classes we have. Another hypothesis is that classes might be similar to each other, making it difficult for the algorithm to differentiate them. For these reasons, we decided to binarize our aggregate measure of vascular risk and run predictions comparing just two classes at a time, such as 0vs1, 0vs2, 0vs3, 0vs4, 0vs5, and 0vs345. In Figure 4.11 we can see the distribution of our binary targets. As a result, we are not predicting and comparing patients with 0, 1, 2, 3, 4, and 5 VRFs at the same time anymore, but instead, we are just comparing two classes at a time. As we can see, these binary targets are still very imbalanced.

Therefore, we applied several techniques to balance them which were described in section 3: Random Oversampling, Random Undersampling, and SMOTE-Tomek. We achieved better accuracies and predictions this time. In addition, we applied all these different techniques to make sure the sampling methods were not affecting our results, and indeed they were not since we saw similar trends and results with all these techniques. For this reason, since the best performing one was SMOTE-TOMEK, in this section we just include the results using this technique. The results with the other sampling techniques can be seen in the Supplementary Data

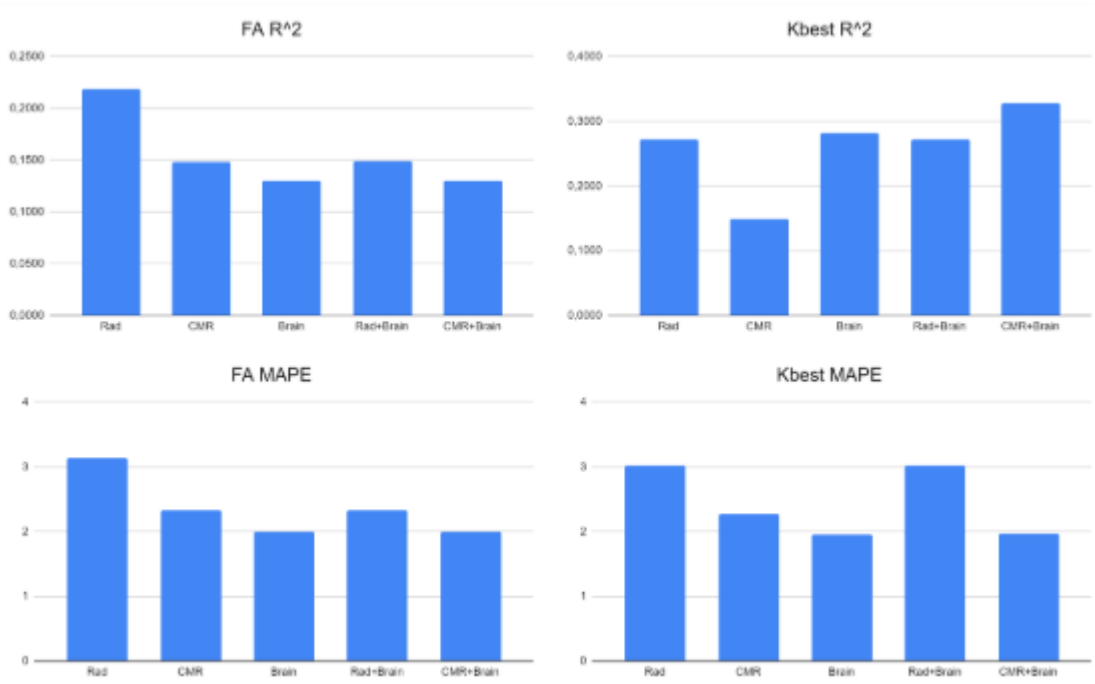

FIGURE 4.7: R^2 and MAPE for gVRF with FA and SelectKbest

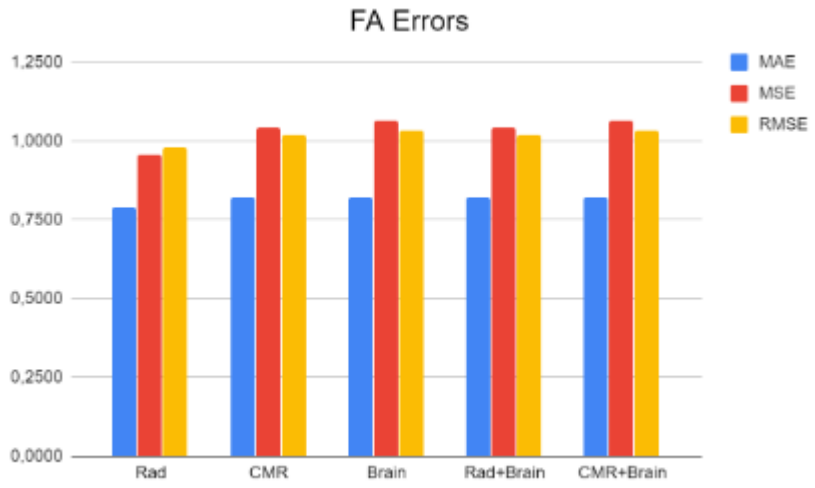

FIGURE 4.8: MAE, MSE and RMSE with FA for gVRF

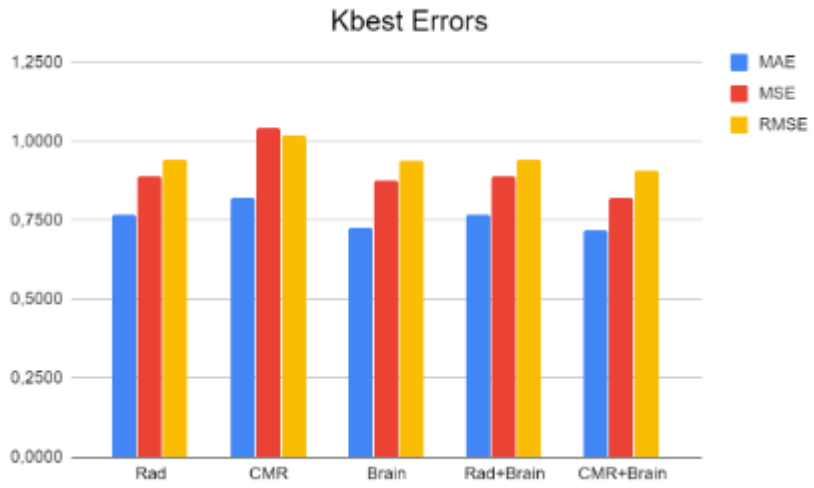

FIGURE 4.9: MAE, MSE and RMSE with SelectKbest for gVRF

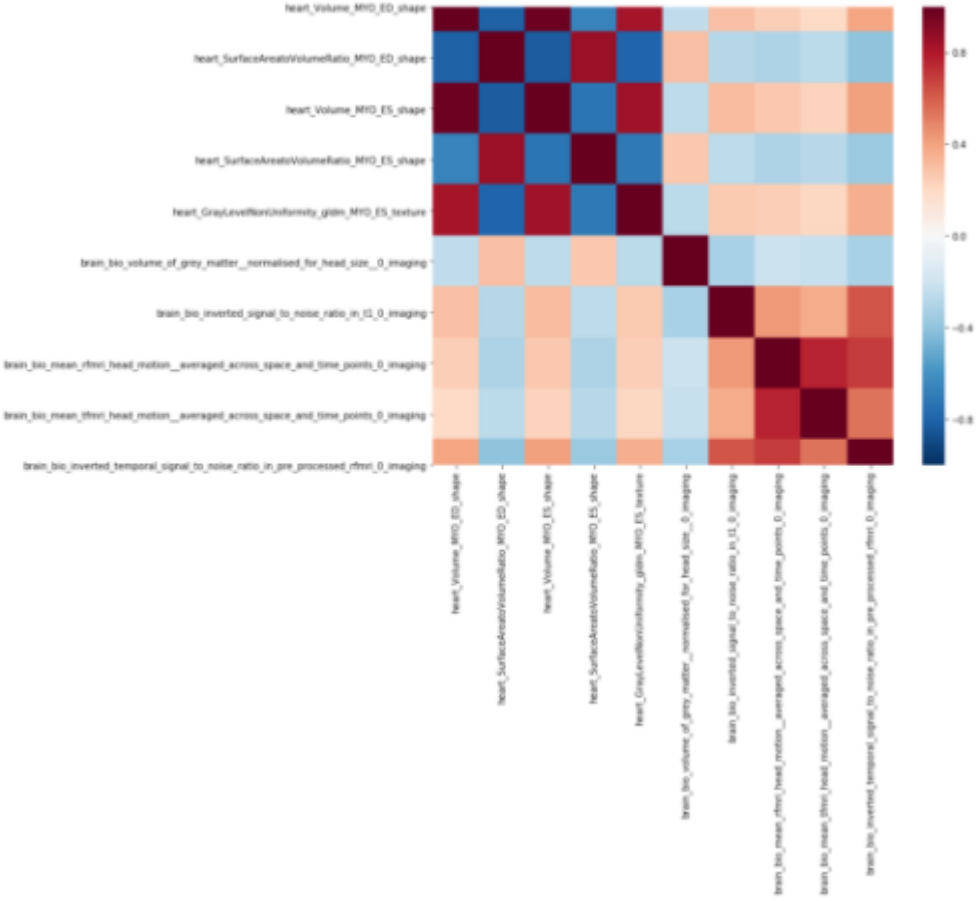

FIGURE 4.10: Correlation plot between the top five SelectKbest selected features from the heart and the brain predicting gVRF

appendix. We obtained the following results when predicting our aggregate measure of vascular risk with random forest and SMOTE-TOMEK:

- In Table 4.7 we can see the accuracy results for the random forest when using FA as the dimensionality reduction technique (we selected five latent factors for each group of variables).
- In Table 4.8 we can see the accuracy results for the random forest when using SelectKbest as the dimensionality reduction technique (we selected the top five SelectKbest features for each group of variables).
- In Figure 4.12 we can see the plots for the accuracy metrics for both dimensionality reduction techniques.
- In Table 4.9 we can see the AUC results for the random forest when using FA as the dimensionality reduction technique.
- In Table 4.10 we can see the AUC results for the random forest when using SelectKbest as the dimensionality reduction technique.
- In Figure 4.13 we can see the plots for the AUC metrics for both dimensionality reduction techniques.
- In Figure 4.14 we can see the plots for the confusion matrix and the ROC Curve for the combination of heart radiomics and brain MRI indices in the case of the 0vs5 binary target.

As we can see in these plots, the main trend is that model accuracy and AUC improves as VRFs burden increases. Therefore, the best results are obtained when this burden is maximized, which is with the comparison between patients with an aggregate measure of zero and patients with an aggregate measure of five (0vs5). These results reinforce the hypothesis mentioned earlier that classes might be similar to each other, especially those with few VRFs (0, 1, 2, and 3). Patients with zero VRFs, the most healthy ones, and patients with five VRFs, the most at risk, seemed to be the most differentiated ones. Furthermore, overall, combined effects of risk factors seem to be better detected by heart CMR radiomics when using FA as the dimensionality reduction technique, while on the other hand, when using SelectKBest as the dimensionality reduction technique, combined effects of risk factors seem to be better detected by brain MRI indices.

Lastly, again as it happened when predicting gVRF, the combination of heart and brain datasets does not seem to improve at all or improve very little our performance metrics, which may lead us to reject our initial hypothesis. However, Figure 4.15 shows the average correlation between the top five SelectKbest features from each dataset. Correlation seems to increase within radiomics features as VRFs burden increases. On the other hand, for brain MRI indices, correlation reaches its highest for the 0vs2 and 0vs3 comparisons, while it stays within the range (0.2, 0.4) for the rest of the binary targets. In the case of heart and brain combinations, correlation stays constant and below 0.2 for all the target classes. Therefore, the average correlation between the top five SelectKbest features from each dataset seems to be very low and not significant, so we cannot infer yet that these datasets do not provide unique information.

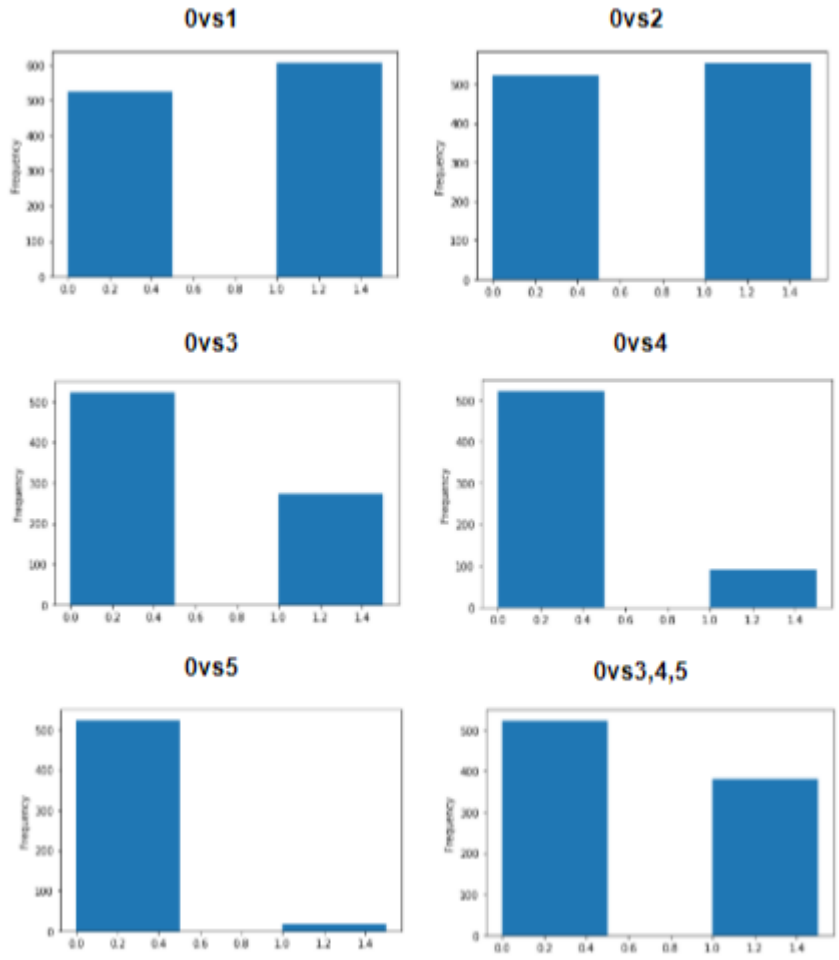

FIGURE 4.11: Binary aggregate measures of vascular risk

TABLE 4.7: Accuracy results predicting aggregate measure with FA

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5482	0.7431	0.7952	0.9058	0.9837	0.7818
Cardio CMR	0.6174	0.6479	0.75	0.8506	0.9426	0.7347
Brain MRI Indices	0.5813	0.6254	0.7694	0.8514	0.9358	0.6981
Heart/Brain Combination	0.5365	0.6525	0.7594	0.8874	0.9640	0.7142
Cardio CMR/Brain Combination	0.5445	0.6692	0.7750	0.8802	0.9610	0.7777

TABLE 4.8: Accuracy results predicting aggregate measure with SelectKbest

Variables	0vs1	0vs2	0vs3	0vs4	0vs5	0vs345
Heart Radiomics	0.5931	0.6937	0.8020	0.7934	0.9675	0.7942
Cardio CMR	0.6006	0.6367	0.7473	0.8316	0.9426	0.7408
Brain MRI Indices	0.6491	0.7269	0.8235	0.8807	0.9710	0.8021
Heart/Brain Combination	0.6563	0.7799	0.8108	0.9147	0.9740	0.8278
Cardio CMR/Brain Combination	0.6589	0.7575	0.8129	0.8947	0.9740	0.8057

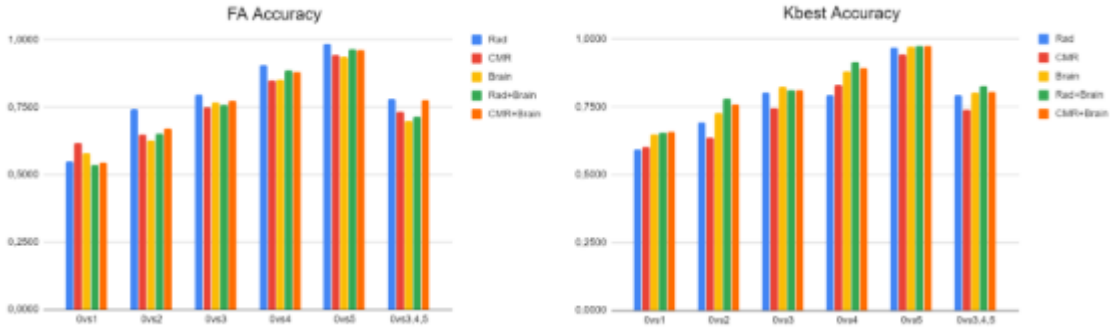

FIGURE 4.12: Accuracy for FA and SelectKbest and SMOTE-TOMEK

TABLE 4.9: AUC results predicting aggregate measure with FA

Variables	0vs1	0vs2	0vs3	0vs4	0vs5	0vs345
Heart Radiomics	0.6013	0.7868	0.8527	0.9645	0.9948	0.8620
Cardio CMR	0.6613	0.7088	0.8435	0.9194	0.9841	0.7801
Brain MRI Indices	0.6113	0.6611	0.8550	0.9318	0.9834	0.7771
Heart/Brain Combination	0.5694	0.6923	0.8204	0.9633	0.9927	0.7582
Cardio CMR/Brain Combination	0.5676	0.6972	0.8509	0.9419	0.9960	0.8574

TABLE 4.10: AUC results predicting aggregate measure with SelectKbest

Variables	0vs1	0vs2	0vs3	0vs4	0vs5	0vs345
Heart Radiomics	0.6198	0.7546	0.8615	0.8660	0.9883	0.8560
Cardio CMR	0.6289	0.6910	0.8275	0.9281	0.9774	0.8278
Brain MRI Indices	0.7327	0.7805	0.8913	0.9555	0.9962	0.8838
Heart/Brain Combination	0.7456	0.8443	0.9098	0.9654	0.9932	0.9198
Cardio CMR/Brain Combination	0.7469	0.8229	0.8999	0.9587	0.9988	0.8872

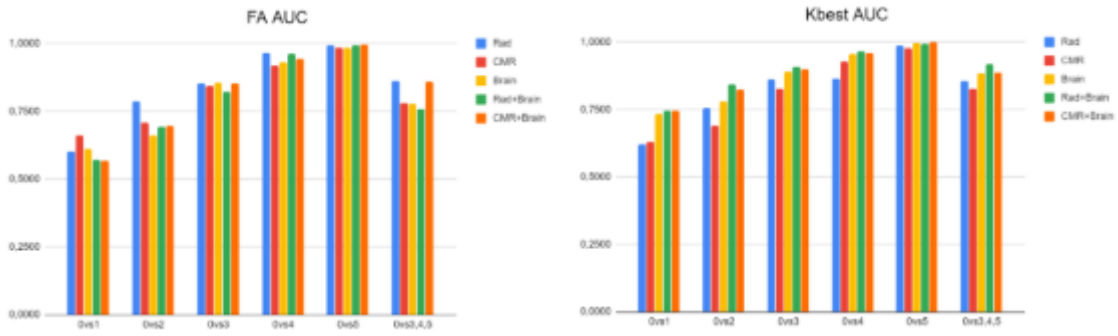

FIGURE 4.13: AUC for FA and SelectKbest and SMOTE-TOMEK

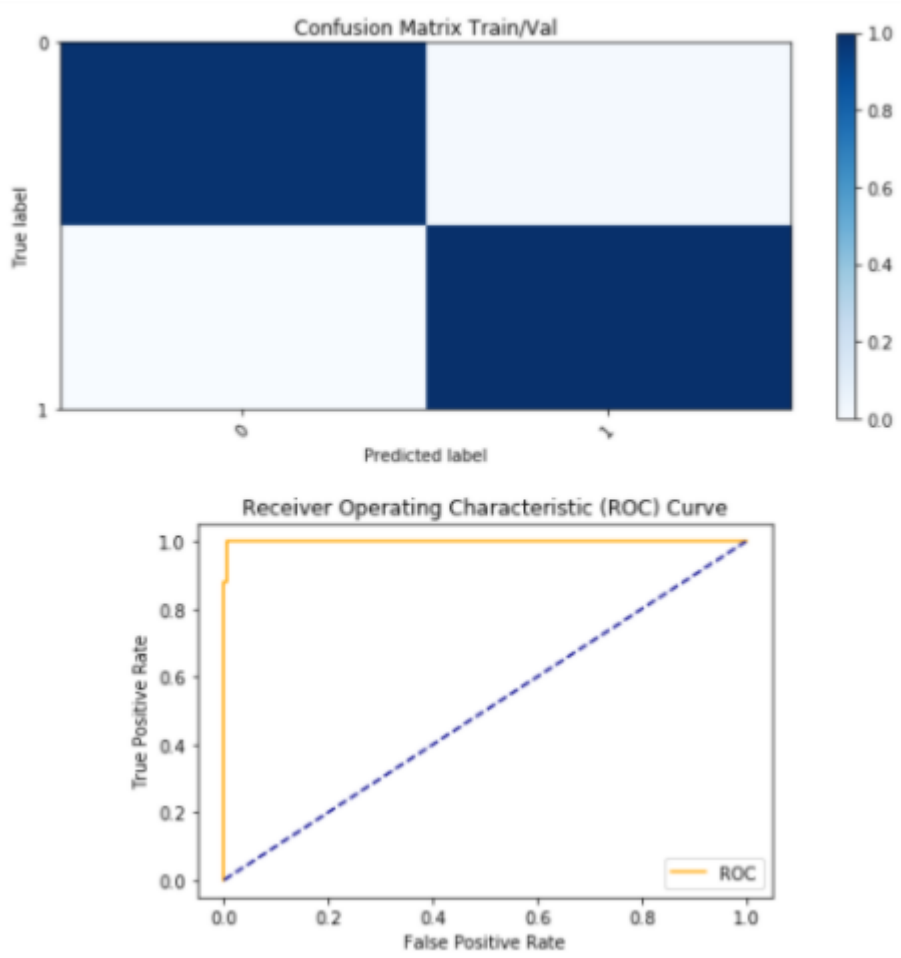

FIGURE 4.14: Confusion matrix and ROC curve for the heart/brain combination predicting the 0vs5 binary target with SMOTE-Tomek.

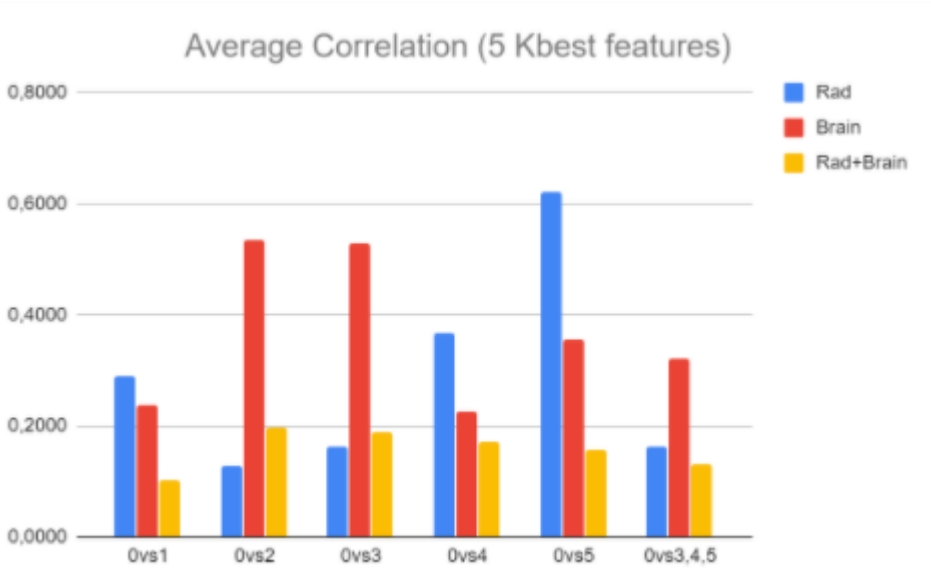

FIGURE 4.15: Average correlation between the top five SelectKbest features from the heart and brain predicting our aggregate measure with SMOTE-Tomek

4.4.3 K-means clustering

From the machine learning experiments we just saw, a new hypothesis emerged, which is that the different classes in our aggregate measure might be increasingly different as the number of vascular risk factors increases. Therefore, making the multi-class classification problem very difficult to differentiate them. The most differentiated classes appeared to be patients with zero VRFs, the most healthy ones, and patients with five VRFs, the most at risk, because when comparing them we achieved the highest performance metrics. However, these hypotheses need to be tested. For this reason, we decided to extract five latent factors from each class with all the features included, the same number of factors we extracted for the ML experiments, and then run the k-means clustering algorithm from the Scikit-learn package in python to find the centroids for each factor in each class. This algorithm clusters data by trying to separate samples in n groups of equal variance. For this task, the k-means algorithm divides a set of samples into disjoint clusters, each described by the mean of the samples in the cluster. These means are known as the cluster "centroids" (Pedregosa et al., 2011).

The results can be seen in Figure 4.16. To be able to generate this plot to visualize the distances between our classes, we just selected the first two centroids derived from the first two latent factors from each class to be able to plot them in a two-dimensional graph. In this plot we can see how S5, patients with an aggregate score of five, and S0, patients with an aggregate score of zero, are the most distant ones, reinforcing our hypothesis that these two classes differ the most. Also, the second largest distance is between S0 and S4, which shows why the comparison between these two aggregate measures achieved the second-highest performance metrics after 0vs5. Lastly, S1, S2, and S3 are the closest aggregate measures to S0, showing why these comparisons obtained the lowest performance metrics and why they are the most similar classes from our target variable.

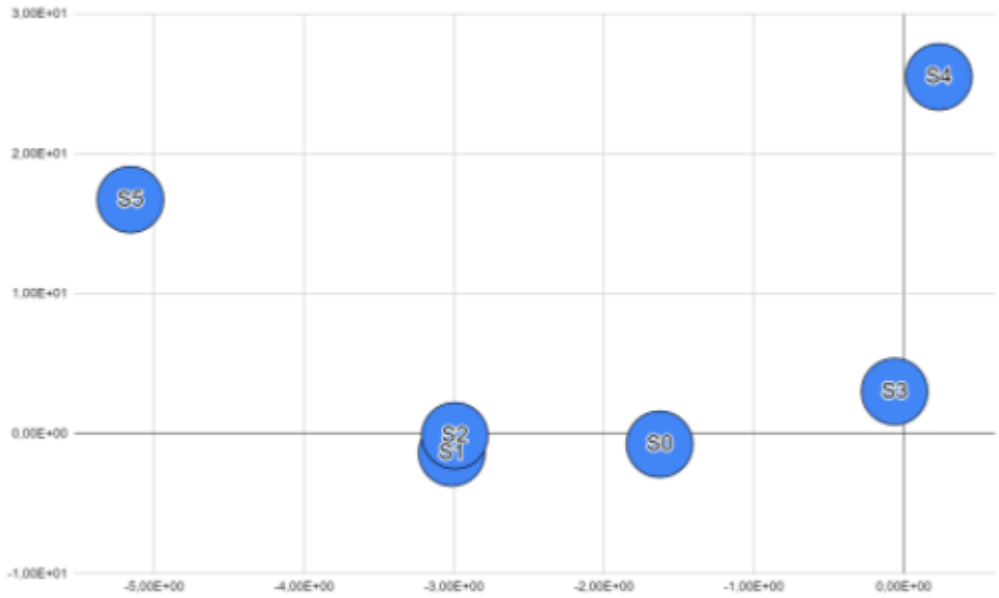

FIGURE 4.16: Centroids derived for each class with the K-means clustering algorithm

4.5 Propensity Score Matching

Individuals with a total of four or five VRFs were each matched on sex and age to a single participant who had no VRFs. First, it is worth mentioning that there is a significant imbalance in our data, as we already explained in the previous sections. The majority group, individuals with zero VRFs have many more records than the minority group, individuals with four and five VRFs. For this reason, we sampled from the majority group when fitting the logistic regression models so that the groups are of equal size. This ensures that more of the majority group is contributing to the generation of propensity scores. Also, according to (Miroglio, 2020) the average accuracy of our models is 75.9 percent, suggesting that there's separability within our data and justifying the need for the matching procedure. Figure 4.17 demonstrates the separability present in our data since test profiles (patients with four and five VRFs) have a much higher propensity, or estimated probability of defaulting than the control group (patients with zero VRFs), given the features we isolated in the data. By default, matches are found from the majority group for the minority group. Therefore, we matched one record from the majority group to each record in the minority group. This was done with replacement, meaning a single majority record could be matched to multiple minority records.

To determine if our data was balanced after the matching procedure we ran some statistical tests to detect any differences between the covariates of our matched test and control groups. First, we looked at Empirical Cumulative Distribution Functions (ECDF) for our test and control groups before and after matching. The results can be seen in Figure 4.18 and Figure 4.19. As we can see, both lines are very close to each other for both of our covariates (sex and age), and even indistinguishable after matching. Also, some other tests and metrics are included in these plots:

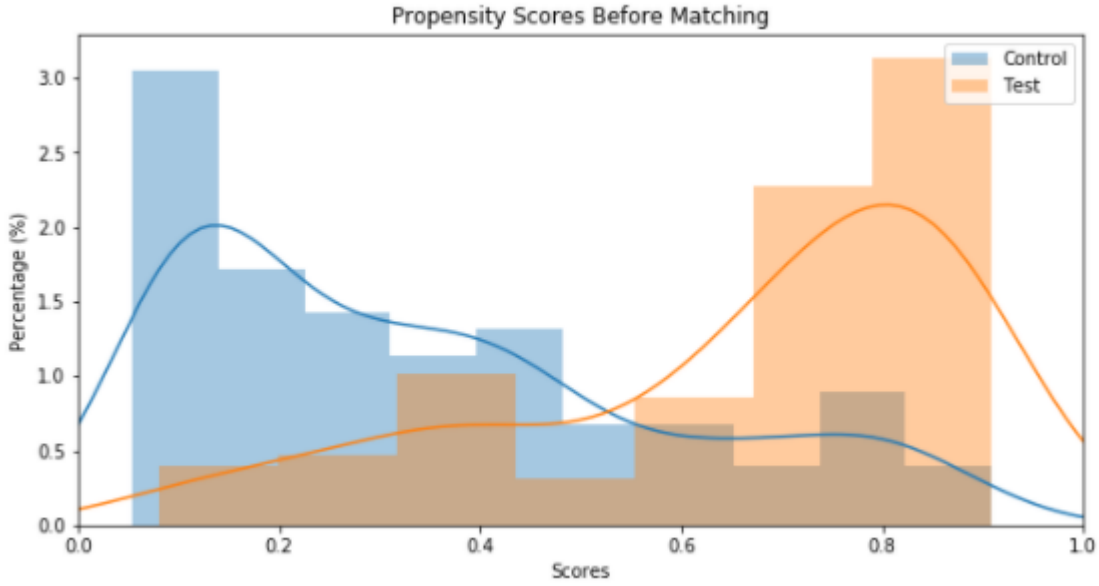

FIGURE 4.17: Propensity scores before matching

- Kolmogorov-Smirnov Goodness of fit Test (KS-test). This test statistic is calculated on 1000 permuted samples of the data, generating an empirical p-value.
- Chi-Square Distance. Similarly, this distance metric is calculated on 1000 permuted samples.
- Standardized mean and median differences.

As we can see the p values from both the KS-test and the grouped permutation of the Chi-Square distance after matching are above 0.05, meaning they were statistically significant, and the standardized mean and median differences are 0.

Lastly, to make sure the matching procedure worked, we also plotted the mean distributions for age and sex before and after matching and computed the means for these covariates for both our control and test groups as we can see in Figure 4.20 and Table 4.11. These results show us how the matching procedure worked since these distributions are more balanced after applying propensity score matching. In addition, it is also noticeable that before matching, there are more males than females in the test group (patients with four and five VRFs). In our data, this sex variable is a binary variable, which is integer encoded with female:0 and male:1. Therefore, the sex mean of 0.75 for the test group before matching proves this claim. Furthermore, the age mean for the test group (59.58) before matching is much higher than the control group (53.79). These results may indicate that age and sex are playing an important role in patients with a tendency to develop VRFs, since as we saw males and older patients have a higher probability of being included in the test group than in the control group. Therefore, although further analysis will be needed, there is a possibility that these confounding variables, age and sex, are affecting the results, and may have a big impact in the study of the heart-brain link.

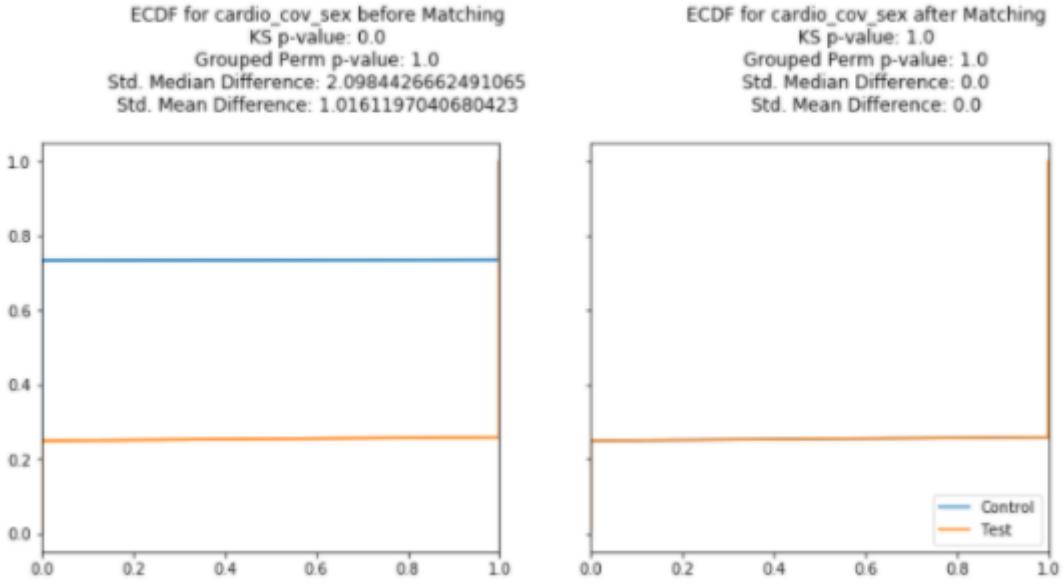

FIGURE 4.18: ECDF for the sex covariate for our test and control groups before and after matching

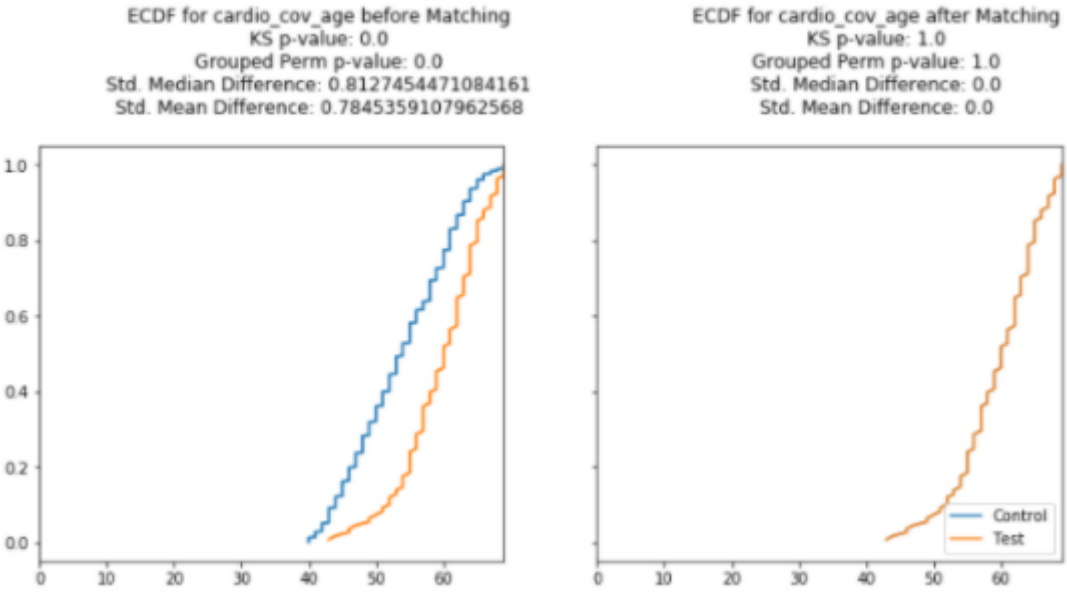

FIGURE 4.19: ECDF for the age covariate for our test and control groups before and after matching

TABLE 4.11: Sex and age means before and after matching

Group	Sex (Before)	Age (Before)	Sex (After)	Age (After)
Test	0.75	59.58	0.75	59.58
Control	0.26	53.79	0.75	59.58

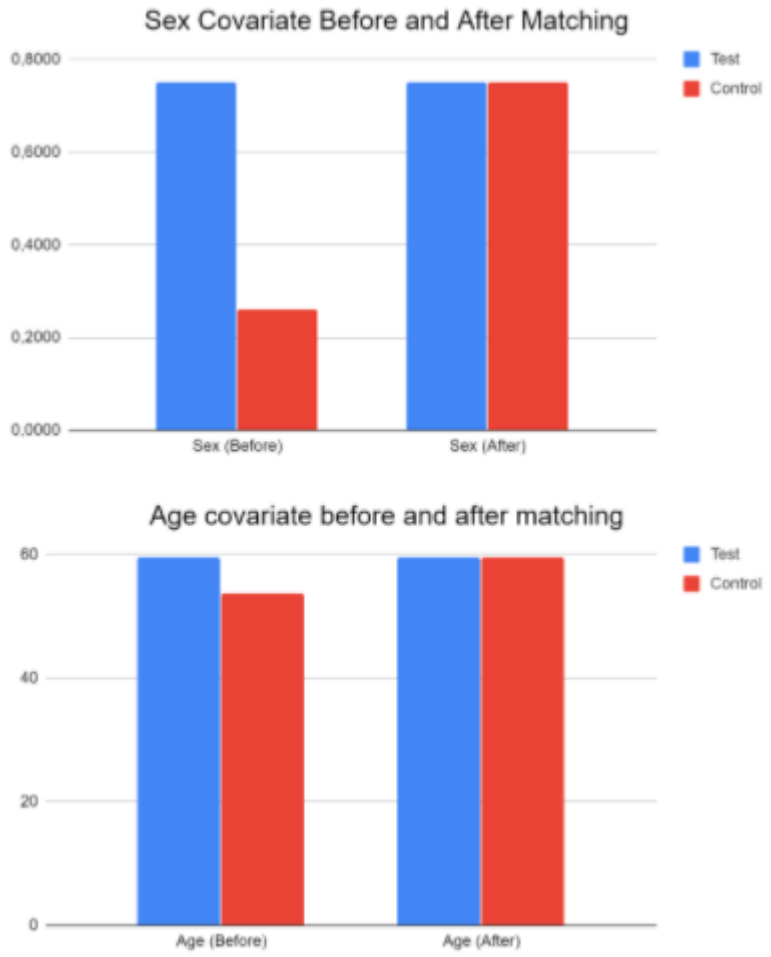

FIGURE 4.20: Sex and age mean distributions before and after matching

4.6 Causal Inference

As we saw with our ML experiments, overall, the combination of brain and heart structure' datasets did not improve at all, or improve very little our performance metrics when trying to predict vascular health, which lead us to reject our initial hypothesis. Therefore, we can infer that these datasets do not provide unique information, and somehow they are related. Also, the combined effects of risk factors were better predicted by heart CMR radiomics when using FA as the dimensionality reduction technique. On the other hand, when using SelectKBest as the dimensionality reduction technique, combined effects of risk factors were better detected by brain MRI indices. These results may suggest that there is a direct link between VRFs and brain structure, as well as a direct link between VRFs and heart structure. However, there is also the possibility that an intermediate factor is playing an important role between these three datasets.

To test these hypotheses we ran mediation analysis with different combinations of graphs with the three datasets. To do so, first, we studied the mediating role that heart structure plays between cardiovascular risk and brain structure. In other words, how much of the connection between cardiovascular risk and brain structure can be explained by changes in heart structure. Secondly, we studied the mediating role that brain structure plays between cardiovascular risk and heart structure. Lastly, we studied the mediating role that cardiovascular risk plays between brain and heart structures. We obtained the following results:

- In Table 4.12 we can see the results from the graphs using our gVRF as the measure of vascular health, and one derived latent factor from both heart CMR radiomics and brain MRI indices.
- In Table 4.13 we can see the results from the graphs using our aggregate measure of vascular risk as the measure of vascular health, and one derived latent factor from both heart CMR radiomics and brain MRI indices.
- In Figure 4.21 we can see the plot with a summary of the results when using our gVRF as the measure of vascular health.
- In Figure 4.22 we can see the plot with a summary of the results when using our aggregate measure of vascular risk as the measure of vascular health.

Table 4.12 and Table 4.13 show the main metrics used to assess the performance of mediation analysis (ACME, ADE, and TE). They also show the order of the datasets in the graphs being tested. The variables in the middle play the mediator role between the first and last variables in each row. As we can see in these tables and plots, we obtained similar trends of results with both our gVRF and our aggregate measure of vascular risk. Also, the most significant results were obtained with the heart structure and the brain structure playing the mediator's roles because the largest, although negative, results for ADE and TE were obtained with the sequences of 'gVRF-Heart-Brain' and 'gVRF-Brain-Heart'. Lastly, in the case of the gVRF or the aggregate measure of vascular risk, sequences 'Heart-gVRF-Brain', and 'Brain-gVRF-Heart', their role as mediators was not significant since their ADE and TE were very low.

For that reason, in Figure 4.23 we have included a summary with the coefficients and metrics obtained for the graph in which heart structure plays the mediator role.

TABLE 4.12: Mediation analysis results with gVRF

Variables	ACME	ADE	TE
gVRF - Heart - Brain	0.014	-0.202	-0.188
gVRF - Brain - Heart	0.011	-0.259	-0.248
Heart - gVRF - Brain	0.057	-0.053	0.004
Brain - gVRF - Heart	0.056	-0.052	0.004

TABLE 4.13: Mediation analysis results with aggregate measure

Variables	ACME	ADE	TE
agg. measure - Heart - Brain	0.008	-0.134	-0.126
agg. measure - Brain - Heart	0.005	-0.223	-0.219
Heart - agg. measure - Brain	0.041	-0.038	0.003
Brain - agg. measure - Heart	0.04	-0.039	0.001

Similarly, in Figure 4.24, a summary with the coefficients and metrics for the graph in which brain structure plays the mediator role is also included. As we can see in these two graphs, we have included the coefficients for the three linear regressions that comprised a mediation analysis, as well as the main metrics being used to assess the performance (ACME, ADE, and TE). Overall, we can see how VRFs and brain structure are strongly negatively correlated, with a significant negative coefficient of -0.1895. Similarly, VRFs and heart structures are strongly negatively correlated as well with a significant negative coefficient of -0.2485. However, heart structure and brain structure are weakly negatively correlated with coefficients below -0.05. Nonetheless, there is a small mediation role that heart structure is playing (ACME: 0.014) as well as the brain structure (ACME: 0.010), which might explain the connection between these datasets, and why they did not provide unique information or improved performance when trying to predict vascular health. So in conclusion, as we expected, in both graphs, X affects Y, X also affects M, and lastly, M affects Y, but the effect of X on Y still exists. Therefore, M partially mediates between X and Y.

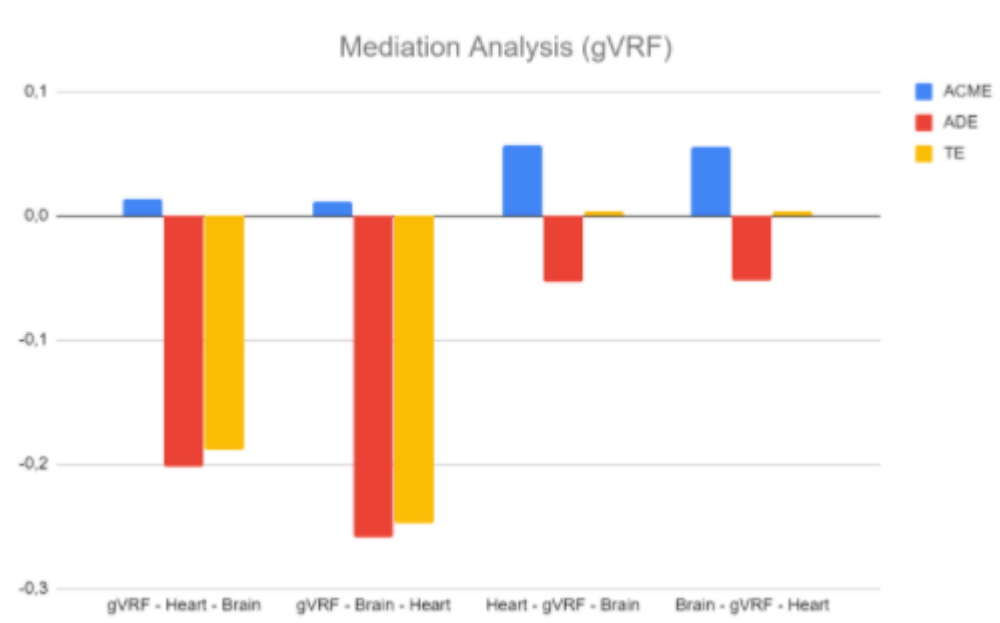

FIGURE 4.21: Results for the mediation analysis with gVRF

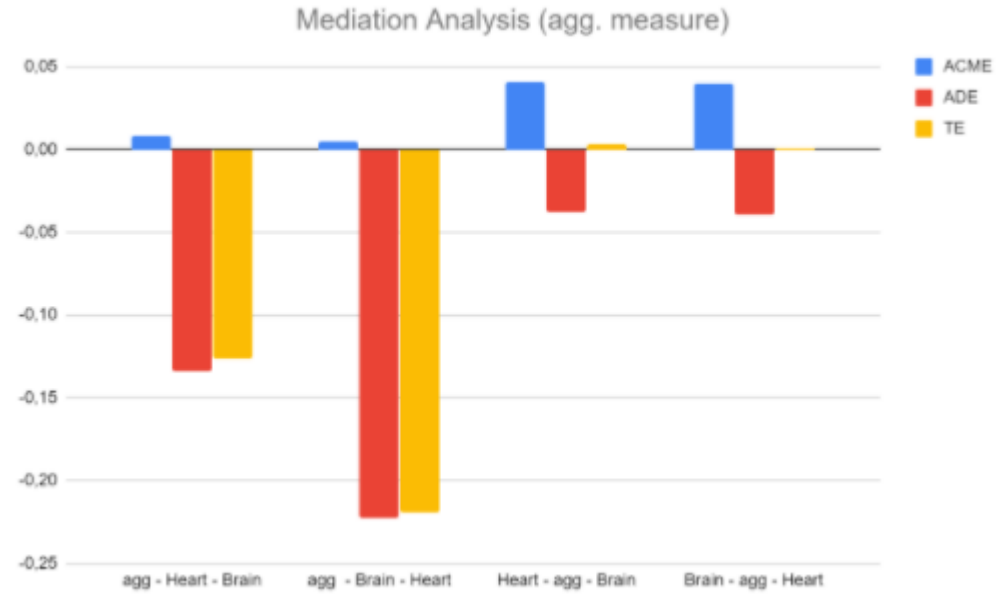

FIGURE 4.22: Results for the mediation analysis with the aggregate measure

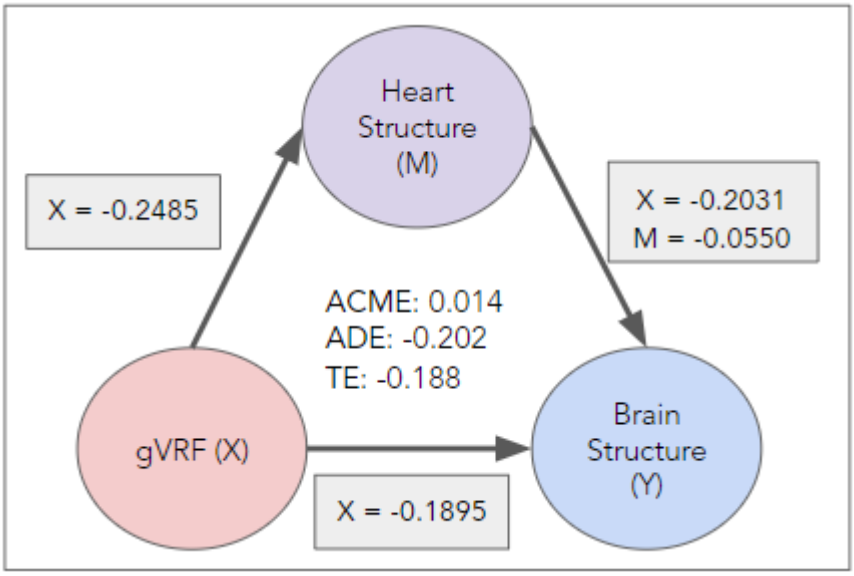

FIGURE 4.23: Graph tested for the mediation analysis with heart structure as mediator (M)

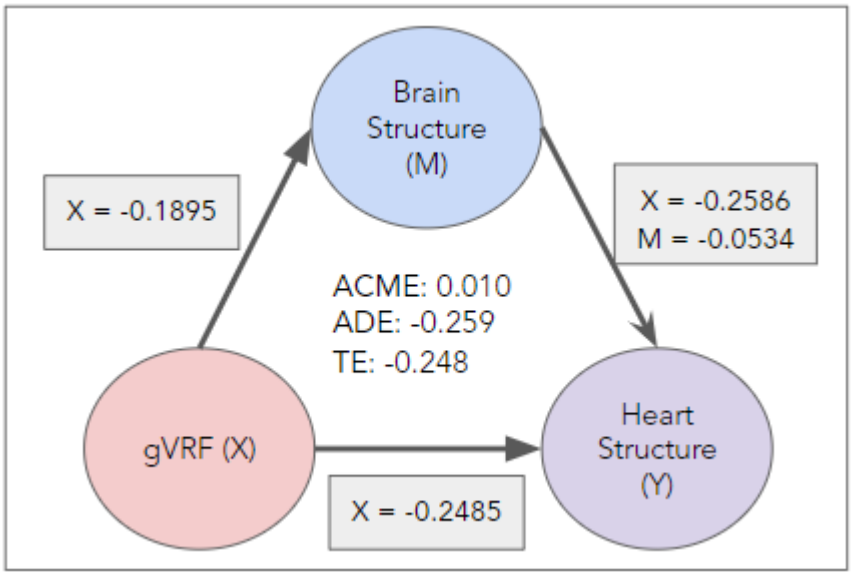

FIGURE 4.24: Graph tested for the mediation analysis with brain structure as mediator (M)

Chapter 5

Discussion and Future Work

In this section, a summary of the main results obtained with the development of this master thesis is presented, as well as a discussion, limitations of the study, and possible future work that can be done.

5.1 Summary of findings and conclusions

In this systemic study we have done an extensive multi-disease research following two different approaches: traditional machine learning techniques and causal inference methods. The traditional machine learning experiments suggested a strong association between vascular health and heart structure, as well as a strong link between vascular health and brain structure. Also, since the combination of brain and heart structure' datasets did not improve the performance of our ML algorithms when trying to predict vascular health, this suggested that another relationship between heart diseases and brain diseases could also exist. Therefore, indicating that the connections between these three factors were not independent, and justifying the need for a simultaneous approach using causal inference techniques.

Causal inference modelling is a large evolving field with many techniques that could be used. However, due to the nature of our data, and the results obtained from the ML experiments, we considered that the best approach was using causal mediation analysis. From this analysis, we confirmed our previous hypothesis and found a strong negative association between vascular health and heart structure, as well as a strong negative association between vascular health and brain structure. Also, although small, another negative association was found between heart and brain structures, which played a significant mediating role in our graphs.

In this thesis, several machine learning and causal inference concepts were implemented to study the associations between heart and brain diseases. To the best of our knowledge, this is the first combined study that explores simultaneously the links between vascular health, heart structure, and brain structure.

5.2 Limitations and future work

In this thesis, for comparison purposes, and due to our limited computational capacity, we only used a subset of features for both traditional and causal experiments. In the case of the traditional ML experiments, vascular health was predicted with a subset of the top five SelectKbest features from the heart and brain structure datasets, as well as a subset of only five latent factors. Furthermore, for the mediation analysis we only used the top factor from the heart and brain datasets respectively. We argue

that with a bigger sample of features and factors involved, the correlation between features from different datasets could increase, the combination of features from different datasets might increase performance, and the coefficients of associations may improve their significance as well.

Also, to measure the vascular health of patients, we computed two different variables following older cohorts: a latent factor of general vascular risk (gVRF), and an aggregate measure of vascular risk. However, as we saw with the k-means clustering algorithm, classes within our aggregate measure of vascular risk were too similar, making the multiclass classification task very difficult to differentiate them. We propose that better ways to measure vascular risk should be implemented to better assess the vascular health of patients in future experiments.

Lastly, as we saw with the propensity score matching analysis, age and sex are two cofounding variables that presented great differences between our control and test groups. Therefore, it would be interesting to expand upon the existing datasets to encompass a more equal and balanced spread of participants in terms of medical, health, and lifestyle factors, that could lead to the avoidance of the effect mentioned above and allow for a more robust study of the associations between heart and brain diseases.

Appendix A

List of Abbreviations

Most common abbreviations used in this thesis and their meaning are collected in this section:

- CMR: Cardiovascular magnetic resonance
- MRI: Magnetic resonance imaging
- BMI: Body mass index
- SBP: Systolic blood pressure
- DBP: Diastolic blood pressure
- WHR: Waist hip ratio
- gVRF: Latent factor of general vascular risk
- KMO: Kaiser-Meyer-Olkin
- FA: Factor analysis
- EFA: Exploratory factor analysis
- CFA: Confirmatory factor analysis
- ML: Machine learning
- MAPE: Mean absolute percentage error
- MAE: Mean of the absolute value of the errors
- MSE: Mean of the squared errors
- RMSE: Square root of the mean of the squared errors
- AUC: Area under the curve
- SMOTE: Synthetic minority oversampling technique
- VRFs: Vascular risk factors
- RF: Random forest
- DTs: Decision trees
- CAD: Coronary artery disease
- HF: Heart failure

- ROI: Region of interest
- ROC: Receiving operating characteristic
- ECDF: Empirical cumulative distribution functions
- ACME: Average causal mediation effect
- ADE: Average direct effect
- TE: Total effect

Appendix B

Supplementary Data

B.0.1 Predicting aggregate measure of vascular risk

Predictions with imbalanced target

We obtained the following results when predicting our aggregate measure of vascular risk with random forest and an imbalanced target:

- Table B.1: FA accuracy results with imbalanced data
- Table B.2: SelectKbest accuracy results with imbalanced data
- Figure B.1: Accuracy for FA and SelectKbest and imbalanced dataset
- Table B.3: FA AUC results with imbalanced data
- Table B.4: SelectKbest AUC results with imbalanced data
- Figure B.2: AUC for FA and SelectKbest and imbalanced dataset

TABLE B.1: Accuracy results predicting aggregate measure with FA and imbalanced data

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5221	0.6728	0.7656	0.8486	0.9506	0.7316
Cardio CMR	0.5103	0.5401	0.6527	0.8270	0.9506	0.5992
Brain MRI Indices	0.5103	0.5648	0.6778	0.8378	0.9506	0.6801
Heart/Brain Combination	0.5014	0.6234	0.6945	0.8108	0.9506	0.6544
Cardio CMR/Brain Combination	0.5103	0.5895	0.7029	0.8540	0.9506	0.6691

TABLE B.2: Accuracy results predicting aggregate measure with SelectKbest and imbalanced data

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5516	0.7345	0.7280	0.8324	0.9567	0.7463
Cardio CMR	0.5132	0.5277	0.6569	0.8378	0.9506	0.5735
Brain MRI Indices	0.6460	0.75	0.7824	0.9027	0.9567	0.7757
Heart/Brain Combination	0.6371	0.7716	0.7991	0.8918	0.9567	0.8382
Cardio CMR/Brain Combination	0.6607	0.7345	0.7782	0.8864	0.9691	0.7830

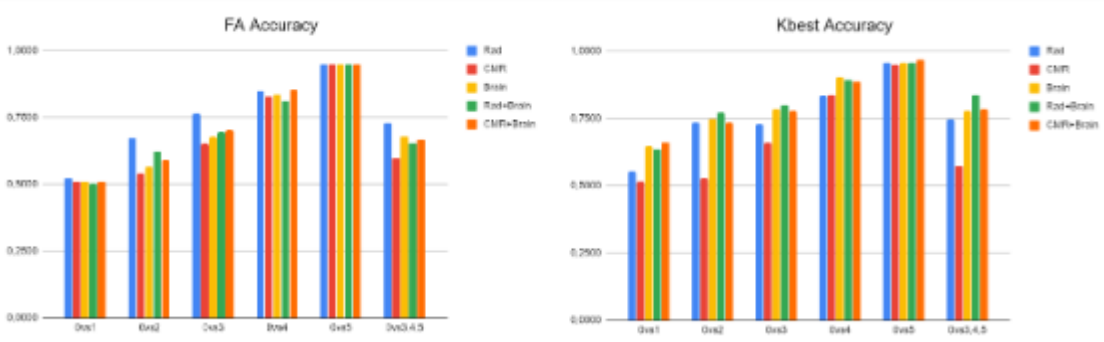

FIGURE B.1: Accuracy for FA and SelectKbest and imbalanced dataset

TABLE B.3: AUC results predicting aggregate measure with FA and imbalanced data

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5128	0.7174	0.7693	0.6816	0.7131	0.7850
Cardio CMR	0.5114	0.5775	0.6112	0.6572	0.5621	0.6176
Brain MRI Indices	0.5339	0.6145	0.6526	0.7590	0.7601	0.6966
Heart/Brain Combination	0.5025	0.6559	0.7077	0.6838	0.5134	0.7053
Cardio CMR/Brain Combination	0.5139	0.6285	0.6747	0.7599	0.7906	0.7077

TABLE B.4: AUC results predicting aggregate measure with SelectKbest and imbalanced data

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5442	0.7689	0.7639	0.7700	0.9058	0.8060
Cardio CMR	0.5300	0.5734	0.6252	0.6198	0.6347	0.5988
Brain MRI Indices	0.6984	0.8198	0.8366	0.9145	0.9152	0.8784
Heart/Brain Combination	0.6850	0.8469	0.8662	0.9065	0.9136	0.9057
Cardio CMR/Brain Combination	0.6967	0.8210	0.8343	0.9104	0.9184	0.8784

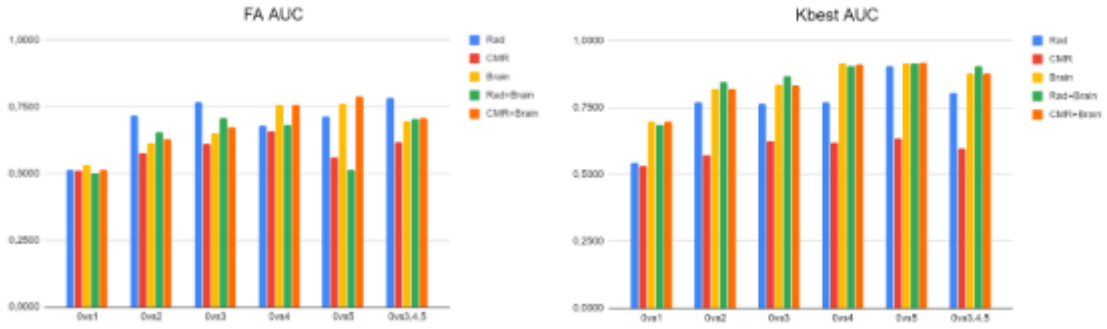

FIGURE B.2: AUC for FA and SelectKbest and imbalanced dataset

TABLE B.5: Accuracy Results predicting aggregate measure with FA and random undersampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5127	0.6592	0.6951	0.6363	0.6363	0.7292
Cardio CMR	0.5063	0.5445	0.6402	0.6727	0.6363	0.6244
Brain MRI Indices	0.5636	0.5573	0.5731	0.7454	0.7272	0.6593
Heart/Brain Combination	0.5286	0.6242	0.7256	0.7272	0.7272	0.6462
Cardio CMR/Brain Combination	0.5700	0.5923	0.5548	0.7090	0.7272	0.6724

Random Undersampling

We obtained the following results when predicting our aggregate measure of vascular risk with random forest and random undersampling:

- Table B.5: FA accuracy results with random undersampling
- Table B.6: SelectKbest accuracy results with random undersampling
- Figure B.3: Accuracy for FA and SelectKbest and random undersampling
- Table B.7: FA AUC results with random undersampling
- Table B.8: SelectKbest AUC results with random undersampling
- Figure B.4: AUC for FA and SelectKbest and random undersampling

TABLE B.6: Accuracy Results predicting aggregate measure with SelectKbest and random undersampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5541	0.6847	0.7256	0.6545	0.6363	0.7336
Cardio CMR	0.5127	0.5796	0.6158	0.6727	0.4545	0.5851
Brain MRI Indices	0.6305	0.7675	0.7926	0.8363	0.7272	0.7554
Heart/Brain Combination	0.6146	0.7611	0.7926	0.8181	0.6363	0.7641
Cardio CMR/Brain Combination	0.6178	0.7420	0.7865	0.8181	0.7272	0.7598

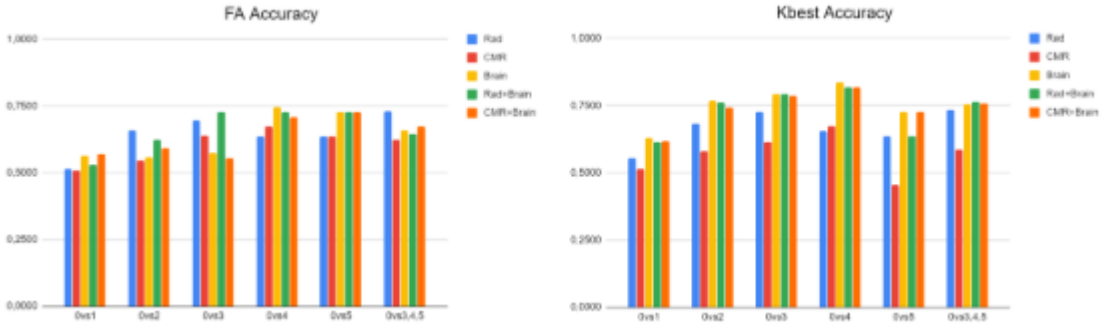

FIGURE B.3: Accuracy for FA and SelectKbest and random undersampling

TABLE B.7: AUC Results predicting aggregate measure with FA and random undersampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5159	0.7080	0.7535	0.7294	0.8667	0.8049
Cardio CMR	0.5251	0.6121	0.6492	0.7659	0.5333	0.6790
Brain MRI Indices	0.5833	0.6117	0.6532	0.8515	0.7167	0.7212
Heart/Brain Combination	0.5072	0.6558	0.7926	0.7759	0.7333	0.7040
Cardio CMR/Brain Combination	0.5913	0.6103	0.6572	0.8216	0.7000	0.7421

TABLE B.8: AUC Results predicting aggregate measure with SelectKbest and random undersampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.5952	0.7181	0.8131	0.7328	0.7667	0.8002
Cardio CMR	0.5234	0.6126	0.6354	0.7142	0.7333	0.6274
Brain MRI Indices	0.6822	0.8375	0.8759	0.9151	0.8833	0.8616
Heart/Brain Combination	0.6421	0.8451	0.8717	0.8886	0.7333	0.8562
Cardio CMR/Brain Combination	0.6758	0.8449	0.8734	0.9271	0.9333	0.8609

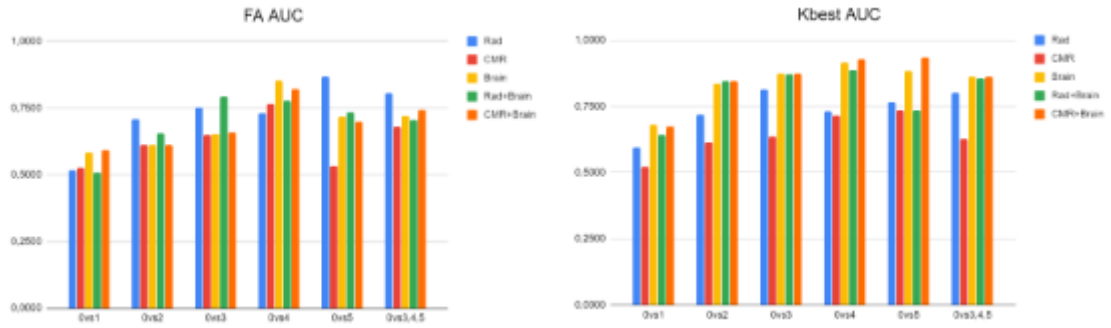

FIGURE B.4: AUC for FA and SelectKbest and random undersampling

TABLE B.9: Accuracy results predicting aggregate measure with FA and random oversampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.6923	0.8048	0.8560	0.9490	0.9936	0.8280
Cardio CMR	0.6950	0.7327	0.7961	0.9203	0.9777	0.7802
Brain MRI Indices	0.7252	0.7057	0.8312	0.9331	1.0	0.8057
Heart/Brain Combination	0.6868	0.7357	0.8089	0.9426	0.9968	0.7707
Cardio CMR/Brain Combination	0.7225	0.7267	0.8312	0.9299	1.0	0.8216

Random Oversampling

We obtained the following results when predicting our aggregate measure of vascular risk with random forest and random oversampling:

- Table B.9: FA accuracy results with random oversampling
- Table B.10: SelectKbest accuracy results with random oversampling
- Figure B.5: Accuracy for FA and SelectKbest and random oversampling
- Table B.11: FA AUC results with random oversampling
- Table B.12: SelectKbest AUC results with random oversampling
- Figure B.6: AUC for FA and SelectKbest and random oversampling

TABLE B.10: Accuracy results predicting aggregate measure with SelectKbest and random oversampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.7142	0.7987	0.8439	0.9458	0.9840	0.8630
Cardio CMR	0.7335	0.7387	0.7866	0.9044	0.9936	0.7515
Brain MRI Indices	0.7582	0.8258	0.8885	0.9585	1.0	0.8949
Heart/Brain Combination	0.7774	0.8138	0.9235	0.9617	0.9936	0.9140
Cardio CMR/Brain Combination	0.7637	0.8138	0.9012	0.9585	1.0	0.9012

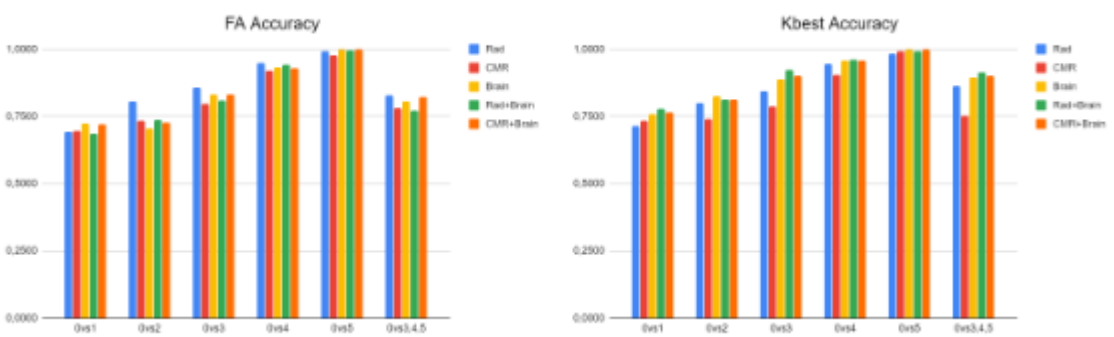

FIGURE B.5: Accuracy for FA and SelectKbest and random oversampling

TABLE B.11: AUC results predicting aggregate measure with FA and random oversampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.7669	0.8602	0.9174	0.9961	1.0000	0.9180
Cardio CMR	0.7958	0.7989	0.8935	0.9934	1.0000	0.8668
Brain MRI Indices	0.7822	0.8155	0.9400	0.9981	1.0000	0.8984
Heart/Brain Combination	0.7662	0.8399	0.9105	0.9956	1.0000	0.8720
Cardio CMR/Brain Combination	0.7868	0.8023	0.9338	0.9953	1.0000	0.8906

TABLE B.12: AUC results predicting aggregate measure with SelectKbest and random oversampling

Variables	**0vs1**	**0vs2**	**0vs3**	**0vs4**	**0vs5**	**0vs345**
Heart Radiomics	0.8160	0.8893	0.9405	0.9953	1.0000	0.9189
Cardio CMR	0.8130	0.8163	0.9069	0.9920	1.0000	0.8638
Brain MRI Indices	0.8492	0.8907	0.9751	0.9975	1.0000	0.9594
Heart/Brain Combination	0.8567	0.9057	0.9810	0.9972	1.0000	0.9649
Cardio CMR/Brain Combination	0.8436	0.8894	0.9736	0.9987	1.0000	0.9600

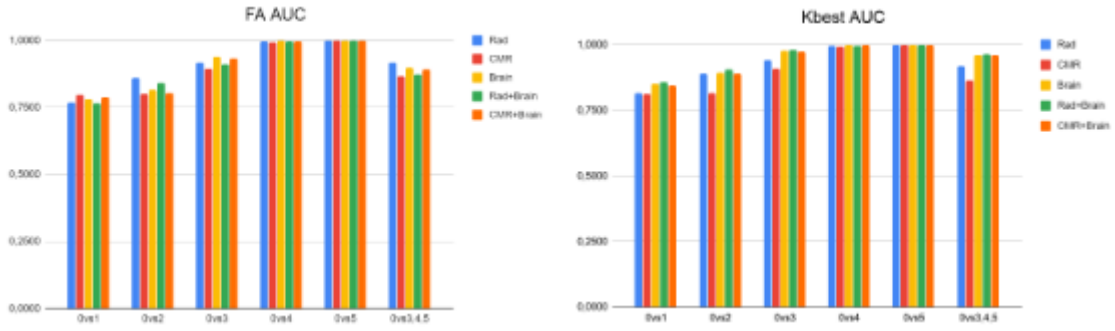

FIGURE B.6: AUC for FA and SelectKbest and random oversampling

Appendix C

GitHub Software Repository

Jupyter notebooks, all python code produced, and support files can be found in the following Github Software Repository.

Bibliography

Alrabghi, Lujain et al. (Aug. 2018). "Stroke types and management". In: *International Journal Of Community Medicine And Public Health*. DOI: 10.18203/2394-6040.ijcmph20183439.

Biobank (2020a). "UK Biobank". In: URL: http://www.ukbiobank.ac.uk/.

— (2020b). "UK Biobank Brain Imaging Documentation". In: URL: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf.

Bruns, Mireia Masias (2017). "Predictive modelling of brain structure and function based on cardiovascular magnetic resonance and radiomics". In:

Cerny, BA and HF Kaiser (1977). "A Study Of A Measure Of Sampling Adequacy For Factor-Analytic Correlation Matrices". In: *Multivariate behavioral research* 12.1, 43—47. ISSN: 0027-3171. DOI: 10.1207/s15327906mbr1201_3. URL: https://doi.org/10.1207/s15327906mbr1201_3.

Cetin, Irem et al. (July 2020). *A radiomics approach to analyze cardiac alterations in hypertension*.

Clark, Michael (2019). *Michael Clark: Mediation Models*. URL: https://m-clark.github.io/posts/2019-03-12-mediation-models/.

Cox, Simon R et al. (Mar. 2019a). "Associations between vascular risk factors and brain MRI indices in UK Biobank". In: *European Heart Journal* 40.28, pp. 2290–2300. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehz100. eprint: https://academic.oup.com/eurheartj/article-pdf/40/28/2290/28968838/ehz100.pdf. URL: https://doi.org/10.1093/eurheartj/ehz100.

Cox, SR et al. (2019b). "Brain imaging correlates of general intelligence in UK Biobank". In: *bioRxiv*. DOI: 10.1101/599472. eprint: https://www.biorxiv.org/content/early/2019/04/04/599472.full.pdf. URL: https://www.biorxiv.org/content/early/2019/04/04/599472.

Dehejia, Rajeev and S. Wahba (Feb. 1998). "Propensity Score Matching Methods for Non-Experimental Causal Studies". In: *Rev Econ Stat* 84.

Dokken, Betsy B. (June 2008). "The pathophysiology of cardiovascular disease and diabetes: Beyond blood pressure and lipids". English (US). In: *Diabetes Spectrum* 21.3, pp. 160–165. ISSN: 1040-9165. DOI: 10.2337/diaspect.21.3.160.

Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak (2016). "Radiomics: Images Are More than Pictures, They Are Data". In: *Radiology* 278.2. PMID: 26579733, pp. 563–577. DOI: 10.1148/radiol.2015151169. eprint: https://doi.org/10.1148/radiol.2015151169. URL: https://doi.org/10.1148/radiol.2015151169.

Gorelick, Philip B. and Farzaneh Sorond (2018). "Vascular risk burden, brain health, and next steps". In: *Neurology* 91.16, pp. 729–730. ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000006346. eprint: https://n.neurology.org/content/91/16/729.full.pdf. URL: https://doi.org/10.1212/WNL.0000000000006346.

Imai, Kosuke, Luke Keele, and Dustin Tingley (Oct. 2010). "A General Approach to Causal Mediation Analysis". In: *Psychological methods* 15, pp. 309–34. DOI: 10.1037/a0020761.

Jefferson, Angela L. et al. (Aug. 2010). "Cardiac index is associated with brain aging: The framingham heart study". English (US). In: *Circulation* 122.7, pp. 690–697. ISSN: 0009-7322. DOI: 10.1161/CIRCULATIONAHA.109.905091.

Lambin, Philippe et al. (Oct. 2017). "Radiomics: The bridge between medical imaging and personalized medicine". In: *Nature Reviews Clinical Oncology* 14. DOI: 10.1038/nrclinonc.2017.141.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

Liu, Peter (July 2008). "Cardiorenal syndrome in heart failure: A cardiologist's perspective". In: *The Canadian journal of cardiology* 24 Suppl B, 25B–9B. DOI: 10.1016/S0828-282X(08)71027-4.

Lyall, Donald M. et al. (Nov. 2016). "Associations between single and multiple cardiometabolic diseases and cognitive abilities in 474 129 UK Biobank participants". In: *European Heart Journal* 38.8, pp. 577–583. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehw528. eprint: https://academic.oup.com/eurheartj/article-pdf/38/8/577/24120983/ehw528.pdf. URL: https://doi.org/10.1093/eurheartj/ehw528.

Martin-Isla, Carlos et al. (2020). "Image-Based Cardiac Diagnosis With Machine Learning: A Review". In: *Frontiers in Cardiovascular Medicine* 7, p. 1. ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00001. URL: https://www.frontiersin.org/article/10.3389/fcvm.2020.0000.

Miroglio, B. (2020). "pymatch's documentation". In: URL: https://github.com/benmiroglio/pymatch.

Ottoboni, K. (2020). "pscore$_{m}$*atch*'*sdocumentation*". In: URL: http://www.kellieottoboni.com/pscore_match/index.html.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pereira, Vitor et al. (Apr. 2012). "Stressed brain, diseased heart: A review on the pathophysiologic mechanisms of neurocardiology". In: *International journal of cardiology* 166. DOI: 10.1016/j.ijcard.2012.03.165.

Petersen, Steffen E. et al. (2013). "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches". In: *Journal of Cardiovascular Magnetic Resonance* 15. ISSN: 1532-429X. DOI: 10.1186/1532-429X-15-46. URL: https://doi.org/10.1186/1532-429X-15-46.

Phloyngam, Naphatthara (2019). "A computational and visualisation tool for investigating associations between cardiac radiomics, risk factors and clinical data". In: URL: https://www.semanticscholar.org/paper/A-computational-and-visualisation-tool-for-between-Phloyngam/3b5235f55ebff65d928a3a1409d2bf387e57f7a4.

Raisi-Estabragh, Zahra et al. (Mar. 2020). "Cardiac magnetic resonance radiomics: basic principles and clinical perspectives". In: *European Heart Journal - Cardiovascular Imaging* 21.4, pp. 349–356. ISSN: 2047-2404. DOI: 10.1093/ehjci/jeaa028. eprint: https://academic.oup.com/ehjcimaging/article-pdf/21/4/349/32932013/jeaa028.pdf. URL: https://doi.org/10.1093/ehjci/jeaa028.

Roger, Véronique L (July 2017). "The heart–brain connection: from evidence to action". In: *European Heart Journal* 38.43, pp. 3229–3231. ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehx387. eprint: https://academic.oup.com/eurheartj/

article-pdf/38/43/3229/21745495/ehx387.pdf. URL: https://doi.org/10.1093/eurheartj/ehx387.

Tahsili-Fahadan, Pouya and Romergryko Geocadin (Feb. 2017). "Heart–Brain Axis: Effects of Neurologic Injury on Cardiovascular Function". In: *Circulation Research* 120, pp. 559–572. DOI: 10.1161/CIRCRESAHA.116.308446.

Tingley, Dustin et al. (Oct. 2014). "Mediation: R Package for Causal Mediation Analysis". In: *Journal of Statistical Software* 59. DOI: 10.18637/jss.v059.i05.

Wardlaw, JM. et al. (2014). "Vascular risk factors, large-artery atheroma, and brain white matter hyperintensities". In: *Neurology* 15, p. 82.

Williams, Brett, Andrys Onsman, and Ted Brown (2010). "Exploratory factor analysis: A five-step guide for novices". In: *Australasian Journal of Paramedicine* 8.3. DOI: 10.33151/ajp.8.3.93. URL: https://ajp.paramedics.org/index.php/ajp/article/view/93.

Yong, An and Sean Pearce (Oct. 2013). "A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis". In: *Tutorials in Quantitative Methods for Psychology* 9, pp. 79–94. DOI: 10.20982/tqmp.09.2.p079.