Universidad del País Vasco
Euskal Herriko Unibertsitatea

University Master's Degree
Computational **E**ngineering and **I**ntelligent **S**ystems

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A
I C S I

informatika fakultatea
facultad de informática

KZAA /CCIA

Master's Thesis

Review of Feature Selection Techniques in Parkinson´s Disease using OCT-imaging data

Reyero Lobo, Paula

Tutor(s)

Inza Cano, Iñaki
Konputazio Zientziak eta Adimen Artifiziala saila
Informatika Fakultatea

Gabilondo Cuéllar, Iñigo
Neurodegenerative Diseases Group
Biocruces Bizkaia-Gurutzetako Unibertsitate Ospitalea

July 2020

Abstract

Background

Several spectral-domain optical coherence tomography studies (OCT) reported a decrease on the macular region of the retina in Parkinson's disease. Yet, the implication of retinal thinning with visual disability is still unclear.

Methods

Macular scans acquired from patients with Parkinson's disease (n = 100) and a control group (n = 248) were used to train several supervised classification models. The goal was to determine the most relevant retinal layers and regions for diagnosis, for which univariate and multivariate filter and wrapper feature selection methods were used. In addition, we evaluated the classification ability of the patient group to assess the applicability of OCT measurements as a biomarker of the disease.

Results

The classification performance results based on mean thickness values of the retinal layers demonstrate high accuracy, precision and recall scores. Furthermore, an improvement of performance is seen including lower cardinality feature subsets, namely, the values the ganglion-cell inner-plexiform layer within the outer macular region, together with the total thickness value of the whole retina.

Conclusion

Several state-of-the-art supervised learning techniques support the hypothesis of retinal thinning as a valid biomarker of the disease, in particular of the ganglion-cell inner-plexiform layer.

Contents

List of Figures

List of Tables

1. CHAPTER

Introduction

1.1 Introduction to the problem

Diagnosis of Parkison's disease at present

One of the greatest challenges facing those who treat Parkinson's disease (PD) is the capacity to obtain an early diagnosis of the condition. PD is a neurodegenerative disorder caused by a pathological dopaminergic deficiency in the basal ganglia of the brain. The sooner the disease is detected, the earlier therapeutic measures can be put in place to delay the progression of neuronal loss.

At present, the diagnosis is still based in a clinical evaluation and there is still no particular test for the detection of the disease, only diagnostic standards which reflect the disease condition. The most widely used standards are the Parkinson's UK Brain Bank or the International Parkinson and Movement Disorder Society criteria, which are specially relevant to begin treatment at early stages.

Meanwhile, a combination of several diagnostic tests are used for this purpose. For instance, the motor score provided by the Unified Parkinson Disease Rating Scale (UPDRS) is a good predictor of Lewy body-associated neuronal loss in the substantia nigra.

Evidence of importance of variables included in this work

Physical manifestations are yet not the only clue to detect PD. Several biomarker researches are trying to become more aware with the symptoms that precede motor manifestations, called the prodomal symptoms. Wollner and Yahr provided the first evidence of

visual system involvement in PD as a reduction in amplitude shown in the electroretinogram [Bodis Wollner and Yahr, 1978]. Visual dysfunction and quantification of retinal thinning in PD has been on wards reported. First to mention, many studies investigate visual processing deficits in PD, such as visual discrimination, visual categorisation and visuospatial orientation skills [Bodis-Wollner, 2013]. Secondly, retinal thinning has been reported since using different measurement optical coherence tomography instruments (OCT) to quantify the retinal thicknesses. These results suggest potential applicability of OCT as a biomarker in PD.

Approach to this problem

In this work several state-of-the-art supervised learning machine learning techniques are applied on retinal thickness data, acquired from a hundred PD patients, to target the hypothesis of retinal thinning as a potential biomarker of the disease. Moreover, an additional goal was to better understand the specificity of retinal degeneration to the disease progression.

1.2 Objectives

In this study, we raise two questions.

- **To which extend are the measurements of retinal thickness values a valid biomarker of PD?** Several supervised classification models are trained on numerical data, corresponding to thicknesses measured by means of a really common imaging technique used mainly in ophthalmology, called Optical Coherence Tomography (OCT). The retinal thicknesses correspond to measurements of the back of the eye (funduscopic images), and yield different parameters according to which layer of the retina they come from, or which region, as the variables used in this work are the result of the average thickness in different sections of the retina.

- **Which are the most relevant retinal layers and regions in the early diagnosis of Parkinson's disease?** This question is addressed from the relevancy point of view. Several feature selection techniques are applied to the data set to output subsets of the most relevant features, e.g. the ones more correlated to the class, which at the same time do not contain redundant information. For this purpose, multivariate filter and wrapper methods are studied using different score metrics. Moreover, rankings

of the most important features are built from a univariate approach to consider the top-k features and the consistency of the results in terms of stability.

1.3 State of the art and related work

Machine learning in clinical research

The purpose of machine learning is to automatically discover patterns, trends and hidden relationships in large amounts of data. Databases in clinical research, similarly to any other field, magnify and powerful analytic tools become indispensable. Machine learning algorithms use statistical techniques that allow computers to extract useful information from a collection of data, construct classifiers which can find complex relationships in the data and validate hypothesis.

Types of medical problems

Diagnosis and prognosis. Clinical research relies in data science techniques to provide solutions in the diagnosis, prognosis, monitoring or quality control processes. Diagnosis procedures use medical signs and symptoms to determine a disease affection. For instance, authors in [Cao et al., 2014] developed a classifier to detect pulmonary nodules in computer tomography images for early-stage lung cancer diagnosis. Prognosis instead aims to predict the likely or expected development of a disease, including whether the signs or symptoms will improve or worsen or stabilise over time and at which pace [1]. In such scenarios, like the one of this work, the search for so-called biomarkers becomes fundamental.

Biomarkers refer to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly [Strimbu and Tavel, 2010]. Biomarkers that have been well characterized and repeatedly shown to correctly predict relevant clinical outcomes across a variety of treatments and populations and such biological measures have become the primary focus in clinical trials. The goal in this work is to integrate machine learning in combination with feature selection to focus on identifying features that can predict a disease versus a control state. This basis has been highly powerful in bioinformatics, for instance, to find biomarkers in various cancer diseases applying these techniques in molecular profiles of tumor sample data [Kourou et al., 2015]. Biometric measures have also been claimed to

[1] https://en.wikipedia.org/wiki/Prognosis

be robust and reliable biomarkers, for instance the measurements acquired by an ECG signal [Pelc et al., 2019]. In the case of measurements from the eyes, specially the ones acquired using the optical coherence technique, less research has been conducted but many publications suggest the importance of the retina due to its close connection with the brain [Chrysou et al., 2019].

Data science provides clinical researches with a variety of models which can predict future behaviors. More specifically, to solve predictive problems (e.g. supervised classification and regression problems), generally known in Artificial Intelligence environments as supervised learning problems. However, data acquisition in this domain is usually restricted by the high cost of clinical trials, or the rareness of the illnesses under examination. These two facts are examples of why the following limitations in the next paragraph need consideration.

Main issues facing clinical researches

To draw conclusions from wide datasets, those characterised by a large number of features (high dimensionality) and small number of instances – common in many domains including the clinical research – feature selection will ultimately offer a subset of the original features, a trained classifier model and an estimate of the classification accuracy. [Kuncheva and Rodríguez, 2018] remarked the need to include feature selection techniques as part of the learning process, to eliminate overly optimistic classification results due to a "peeking" effect caused by using the dataset twice: for selecting the best subset and evaluating performance.

Another really common issue in the medical domain when using machine learning methods to test hypotheses emerges from imbalanced distribution among problem classes. The accuracy reached with the given learning problem is also different for each class. To account for this fact, special consideration should be given to the preprocessing step.

Review on data level pre-processing methods

The main stages involved in any learning process include the selection of goals, the preprocessing of the data (selection, preparation, transformation and/or reduction of the feature data set), the construction of the model and the analysis of the results.

[Fernández et al., 2018] conducted an extensive research on the main approaches to deal with classification learning from imbalanced data in multiple domains, including engineering, business management, security, bioinformatics and medicine. Their work discerned approaches from the data preprocessing-level (resampling and Synthetic Minority

Oversampling Technique) and from the algorithm-level (cost-sensititive methods and algorithm modifications such as ensemble methods). Data sampling methods modify the training instances to produce a more balanced distribution that yields less biased results towards the minority class (the positive class). This can be performed eliminating instances from the majority class (undersampling), replicating instances from the minority class (oversampling) or with a combination of both approaches.

Regarding undersampling, some represantive works in this area include Wilson's edited nearest neighbor (ENN) rule [Guan et al., 2009], in which noisy instances from the training set are removed until all remaining instances have a majority of their neighbors with the same class and which has been further developed in posterior publications to improve performance. However, these techniques may disregard relevant data and be detrimental to the induction process. Therefore, more sophisticated methods have been developed and used, such as the Synthetic Minority Oversampling Technique (SMOTE) [Chawla et al., 2002], which created new minority class examples by interpolating several minority class instances that lie together for oversampling the training set. The main drawback of oversampling methods, concretely of SMOTE techniques, according to firstly claimed by the authors in [Puntumapon and Waiyamai, 2012] is that, even if the synthetic data are less specific to the original data, the over-generalization problem may occur. New data is generated merging two minority data and, as a result, decision regions are larger, i.e. new data closer to the decision boundary is generated. Over-generalization occurs when synthetic data is generated into the majority class region, leading to misclassification of the non-minority class into the minority class.

Review on feature selection techniques

Modern biomedical data require feature selection methods that can be applied to large scale feature spaces (biomedical measurements), function in noisy contexts, detect complex patterns of association, be flexibly adapted to various problem domains (e.g. gene expression, and clinical data) and be computationally feasible. In this context, to find the minimum number of features that can be representative in a classification problem can lead to more efficient and less biased results, avoiding overfitting.

Dimensionality reduction has been long to make classification more efficient and less error-prone. Many feature selection strategies have been proposed over the years, generally falling into one of the three categories: (1) filter methods, (2) wrapper methods, or (3) embedded methods. Filter methods work within an attribute selection independent scheme, as there exists no relation between the selection process and the learning scheme

and the output is a rank of the features more correlated to the class, and less feature interdependent in the case of multivariate filters. The second two are part of an attribute selection specific scheme, as the algorithms are linked to the learning process and yield subsets of the best features. FS has been proved to be indispensable in any process learning from data, and across different disciplines [Jovic et al., 2015]. This is the case in this study, as the data size was limited. Many FS techniques were tested to avoid the problem of overfitting, i.e. a problem emerging when a combination of features that discriminates purely by chance the class variable due to an over-proportion of the number of features in respect to the number of instances can occur.

2. CHAPTER

Dataset description

2.1 Participants and Study Design

Recruitment and diagnostic categories

The data set is part of a cross-sectional study of patients with Lewy Body Diseases (LBDs), which encompass a spectrum of disorders pathologically characterized by the widespread deposition of Lewy bodies in the central nervous system. 100 patients were recruited for evaluation, including idiopathic's Parkinson's Disesase (iPD, $n = 78$), Dementia with Lewy Bodies (DLB, $n = 7$) and certain genetic variants of Parkinson's Disease (PD), such as the E46K mutation in the α-synuclein gene (E46K-*SNCA*, $n = 8$), the Parkin gene (PARK2, $n = 4$) and the Leucine-Rich Repeat Kinase 2 gene (LRRK2, $n = 3$), and 248 controls. E46K-*SNCA* is recognized as the most pathogenic mutation inducing PD [Zarranz JJ, 2004].

Participants were recruited through the Department of Neurology at Cruces University Hospital and the Biscay PD Association (ASPARBI). All subjects were participants from another study which counted with 37 controls [Murueta-Goyena et al., 2019]. The rest of the control participants were included meeting the same criteria as in that previous study, to have extra OCT scans.

Medical protocol

Controls were selected to approximately match E46K-*SNCA* carriers in age and sex. Only controls with a maximum of minus/plus 6.00 dioptres ($\pm D$) were recruited for this study.

Patients with iPD fulfilled Parkinson's UK Brain Bank criteria for the diagnosis of probable Dementia with Lewy Bodies (DLB) by 2016 revised criteria for the clinical diagnosis of DLB. All patients were studied in an on-medication condition to complete all study assessments.

The study protocol was approved by the regional Basque Clinical Research Ethics Committee. All participants gave written informed consent prior to their participation in the study, in accordance with the tenets of the Declaration of Helsinki.

2.2 Clinical Evaluation

Before presenting the nature of the variables used in this study, it is important to evaluate the phenotype of the patient class. For this purpose, the following demographic variables and clinical tests were recorded. The average values and standard deviation of each variable mentioned on the following are presented in Table 2.1.

Clinical evaluation

Age and sex were recorded for all participants, and disease duration and age of disease onset of patients. Two neurologists experienced in the field of movement disorders recorded Hoehn & Yahr Scale score (HY), Unified Parkinson's Disease Rating Scale scores (UPDRS): part I (UPDRSI)-mentation, behaviour and mood, part II (UPDRSII)-activities of daily living, part III (UPDRSIII)-motor examination and part IV (UPDRSIV)-complications of therapy; Levodopa Equivalent Daily Dose (LEDD) and the Montreal Cognitive Assessment scale (MoCa).

All these rating scales aim to assess the symptoms of the condition and quality of life of the patients. Table 2.1 shows the mean and standard deviation values outlined above of all controls and patients, as well as for each diagnostic category group.

Demographics

The age of disease onset is expressed as a decimal number in years. Disease duration is also expressed in years, calculated from the difference between the evaluation (ev) and diagnosis ($diag$) dates in day units divided by 365:

$$\frac{DD_{ev}/MM_{ev}/YY_{ev} - DD_{diag}/MM_{diag}/YY_{diag}}{365} \tag{2.1}$$

Variables, units	**Control**	**All patients**	**iPD**	**DLB**	**E46K-SNCA**	**PARK2**	**LRRK2**
n	248	100	78	7	8	4	3
Age, years	55.73 (11.8)	61.63 (11.02)	62.68 (8.92)	75.11 (6.48)	48.44 (11.78)	49.9 (14.2)	53.8 (9.1)
Males, n (%)	97 (39.11)	66 (66)	51 (65.38)	5 (71.43)	5 (62.5)	3 (75)	2 (66.67)
Disease duration, years	NA	6.13 (4.61)	5.83 (4.11)	8.47 (5.6)	6.63 (4.79)	8(8.72)	5.07 (3.77)
Age of disease onset, years	NA	55.78 (10.66)	56.78 (8.52)	66.67 (9.32)	43 (9.48)	41.82 (18.88)	48.7 (5.41)
LEDD, mg/day	NA	580.26 (388.13)	598.87 (375.45)	539 (322.11)	460.19 (558.93)	380 (190.96)	713 (292.29)
UPDRSI total score	NA	2.23 (1.99)	2 (1.73)	5.17 (2.03)	3.29(2.6)	1 (0.82)	1 (0.82)
UPDRSII total score	NA	11.28 (6.29)	10.91 (5.62)	16.33 (5.93)	10.43 (9.93)	11 (4.97)	13 (8.29)
UPDRSIII total score	NA	25.94 (11.09)	25.76 (9.57)	36.67 (13.68)	19.57 (17.9)	23.33 (8.22)	26.33 (7.32)
UPDRSIV total score	NA	3.73 (3.75)	3.79 (3.8)	2.5(2.57)	3.43 (3.96)	3 (2.83)	6 (3.74)
Hoehn & Yahr score	NA	2 (0.0-4.0)	2 (1.0-4.0)	2.75(2.0-3.0)	1.5 (0.0-3.0)	2 (1.0-3.0)	2 (2.0-2.5)
MoCA total score	27.02 (3.04)	23.61 (4.58)	24.06 (3.51)	16.86(7.2)	23.88 (6.51)	25.75 (2.95)	24 (3.56)

Table 2.1: Demographics and PD characteristics for each diagnostic category.
Results are displayed as the average value (standard deviation) except for sex and Hoehn & Yahr score, which are shown as number of males (% of males) and as median (range), respectively. iPD, idiopathic Parkinson's disease; DLB, dementia with Lewy bodies; E46K-*SNCA*, patients with E46K mutation in α-synuclein (*SNCA*) gene; PARK2, patients with mutation in parkin gene; LRRK2, patients with mutation in leucine-rich repeat kinase 2 gene; LEDD, Levodopa Equivalent Daily Dose; UPDRS, Unified Parkinson's Disease Rating Scale; MoCA, Montreal Cognitive Assessment; NA, nonapplicable.

Neurological Assessment

The Unified Parkinson's Disease Rating Scale (UPDRS) [1] contains elements of several scales to monitor the course of Parkinson's and the degree of disability. Some of the elements evaluated in part I are: intellectual impairment, thought disorder, motivation/initiative or depression. The second part covers speech, salivation, walking, falling, handwriting, among others. The motor evaluation in the part II includes action tremor, rigidity, finger taps, hand movements, posture, gait and postural stability. Finally, the last part encompasses complications of therapy such as dyskinesia, anorexia or sleep disturbance.

Hoehn & Yahr Scale score (HY) [Hoehn and Yahr, 1967] evaluates the progression of Parkinson's symptoms and the level of disability through eight possible conditions. Each condition indicates: no signs of disease (Stage 0), unilateral symptoms (Stage 1), unilateral symptoms involving also neck and spine (Stage 1.5), bilateral symptoms without impairment of balance (Stage 2), mild bilateral symptoms but can maintain balance if pulled from behind (Stage 2.5), balance impairment but mild to moderate - physically independent (Stage 3), severe disability but with ability to walk or stand unassisted (Stage 4), and need of a wheelchair if unassisted (Stage 5).

Levodopa is nowadays an effective and well-tolerated dopamine replacement to treat Parkinson's disease. The Levodopa Equivalent Daily Dose (LEDD) gives an estimate of the total amount of levodopa in *mg* that results out of the contribution of each drug taken

[1] https://es.wikipedia.org/wiki/Unified_Parkinson%27s_Disease_Rating_Scale

daily by the patient.

Cognitive Assessment

Montreal Cognitive Assessment scale (MoCa) [2] was developed in 2005 to assess the degree of affectation on an individual's cognitive function [Z.S. Nasreddine, 2005]. It evaluates five different cognitive domains: memory, visuospatial skills, executive function, attention/concentration/working memory, language and orientation. The maximum possible score is 30, and scores over 26 are considered normal.

2.3 Brief retina anatomical description

Retina anatomy

The main concepts of the retina anatomy are introduced in this section. The retina is the innermost layer of tissue of the eye, which converts the light into chemical and nervous signals, transported to the visual cortex by way of the optic nerve, see Fig. 2.1.

The fovea, marked in the figure, is the specialized region of the human retina that drives the majority of the visual functions and whose variation in shape is related to alterations of the retinal layers [Provis et al., 2005]. It is in the fovea where the highest density of photoreceptors and neural machinery is located, and from which the connection to the primary visual cortex of the brain through the optical nerve is enabled. The points surrounding the foveal depression form the macular ridge and, within, the slope that rises more abruptly conform what is known as the foveal edge. Most studies did not observe a decrease of the center of the foveal pit, but of these surroundings, which constitute the macular region [Chrysou et al., 2019].

The retina is a multilayered tissue depending on different cell types. Looking into the eye from the outside, these layers comprise the peripapillary retinal nerve fiber layer (mRNFL), the ganglion cell layer (GCL), the inner plexiform layer (IPL), the inner nuclear layer (INL) and several outer retinal layers, including outer plexiform (OPL) and outer nuclear (ONL) layers and the photoreceptor layer (phot). The values of thickness of each layer and the sum of all of them, which corresponds to the total retina thickess, are shown in Fig. 2.4 (B).

Clinical applications

[2] https://en.wikipedia.org/wiki/Montreal_Cognitive_Assessment

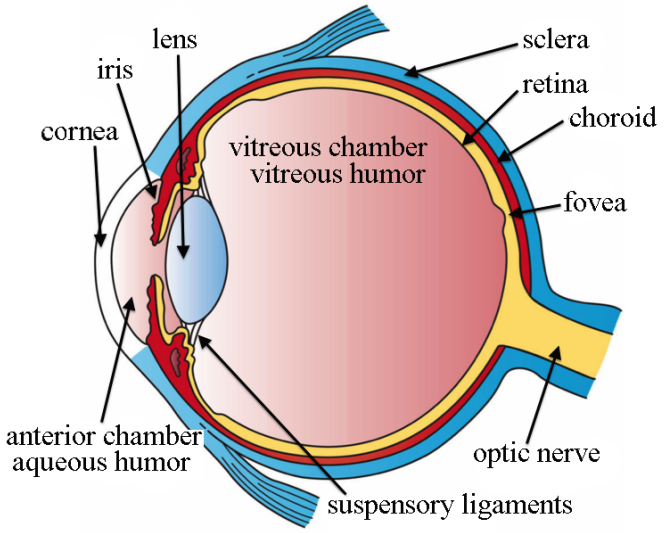


Figure 2.1: Anatomy of the eye.
Location of the retina, fovea and optical nerve. Image source: Wikipedia entry on "The structures of the eye labeled".

The retinal ganglion cell (GC) has been investigated for many purposes, as it is the output neuron of the retina that sends its axon to the brain via the optic nerve. The death and disappearance of GC is especially apparent in the macula, where many GCs are concentrated and where the ganglion cell layer (GCL) is many cell bodies thick. Its thickness becomes relevant, for instance, for reducing variance for glaucoma diagnosis [Knighton et al., 2012], to provide an insight of the development of the foveal pit morphology [Springer AD, 2005] or has been proven to have a strong relationship with the foveal avascular zone (FAZ) [Dubis et al., 2012], the region sorrounding the fovea without retinal blood vessels whose reduction of size has been related with the outspring of several diseases, such as retinopathies (i.e. capillary dropout) in diabetic patients. A recent meta-analysis from 77 studies, totalling 1916 Parkinson's disease patients and 2006 controls, showed significant thinning of the inner retinal layers, resembling changes found in glaucoma and other neurodegenerative diseases like Alzheimer's [Chrysou et al., 2019].

Besides, as the adjacent inner plexiform layer (IPL) becomes often indistinct, their combination (GCIPL) serves as a surrogate for GCL in many clinical purposes. Further investigation of the thickness in this layer becomes of interest, because previous work suggest that in patients with PD it can predict disease severity [Tian et al., 2011, Bodis-Wollner et al., 2014, Jiménez et al., 2014, Garcia-Martin E, 2014, Chrysou et al., 2019, Murueta-Goyena et al., 2019].

All in all, thickness maps can determine the size and shape of the foveal depression and the surrounding macular ridge, two important factors with potential clinical application,

for instance, as potential biomarkers of Parkinson's disease.

Related technology

Spectral-domain Optical Coherence Tomography (OCT) has become the main technique used nowadays to provide an insight of the retinal anatomy, due to being a non-invasive technique that is inexpensive and widely available and only takes up to a few minutes per eye. OCT is an optical signal acquisition and processing method that quantifies differences in optical properties of different layers of the retina. OCT depth information is obtained by a *spectrometer* and analyzed by a *Fast Fourier Transformation*. First, scans of the entire area can be obtained by acquiring successive adjacent B-scans and second, real-time image enhancement is used to reduce the signal to noise ratio (SNR) and improve the definition of a single image by averaging multiple images. This is possible, as the examination is performed simultaneously in various axes, allowing 3D reconstruction [Medical Advisory Secretariat, 2009].

Furthermore, OCT allows for intra-retinal layer segmentation, which is fundamental in the research of the retina as biomarker for PD. Quantifying the layers of the retina raises the hope of developing an in vivo marker for the disease. Nevertheless, as most imaging studies do, OCT yields masses of data, so the definition of *where* (region of interest) and *what* (layer of interest) to focus on must be considered to determine whether retinal thinning parallels disease progression [Hajee et al., 2009].

2.4 OCT Acquisition, Segmentation, and Processing

Data acquisition protocol

Macular volumetric images were obtained using the Spectralis Spectral-Domain OCT-System (Heidelberg Engineering, Heidelberg, Germany).

All macular scans were centered on the fovea. The output consisted of 25 single horizontal axial scans covering a $20^\circ \times 20^\circ$ area (see blue squared region in Fig. 2.2 (B)), with 512 A-scans per B-scan. Each B-scan is acquired as the the average of 49 frames (automatic real-time tracking: 49). Fig. 2.2 (A) shows a B-scan, which consists of a set of, in this case, 512 vertical A-scans.

Data segmentation

The software of the system allows to automatically export the thickness values for each one of the seven layers segmented: RNFL, GCL, IPL, INL, OPL, ONL, Phot. It also

exports the total thickness values of the retina (Retina). Thicknesses are calculated for each A-scan, as the distance length from the upper part of the image to the corresponding layer, measured in μm.

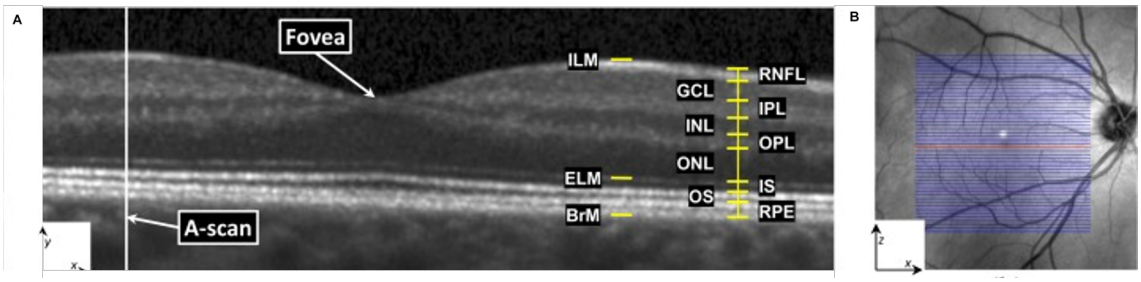


Figure 2.2: Horizontal B-scan with labeled layers on the right-hand side. Every B-scan consists of a set of vertical lines (A-scans) (A). A funduscopic image with lines overlaid representing the locations of every B-scan within a volume (B).
The red line corresponds to the B-scan in (A). Image source: [Lang et al., 2013].

All acquisitions were obtained by the same experienced operator, and the built-in tracking system was used to compensate for eye movements.

All OCT images fulfilled quality control criteria from OSCAR-IB consensus, accounting for Obvious problems (O), poor Signal strength (S), Centration of scan (C), Algorithm failure (A), Retinal pathology other than PDrelated (R), Illumination (I) and Beam placement (B) [Tewarie et al., 2012].

Data Processing

The raw thickness values in each region of the Early Treatment Diabetic Retinopathy Study grid (ETDRS, shown in Fig. 2.3) were used to compute thicknesses in different discs and rings, to enrich the measurements from the anatomical point of view. The variables included in the study are explained in section 2.5. Their values were computed by means of a weighted average value for that region. For example, the average thickness value of a ring in the 1-to 3-mm area region would be the result of the sum of $S3, N3, I3$ and $T3$ divided by 4-according to the figure- whereas the 1-mm disc would be directly the value of $C1$.

2.5 Variable Description

The built-in software of the Spectralis OCT equipment (HRA Spectralis Viewing Module version 6.0.9.0) automatically segmented macular layers, as previously mentioned.

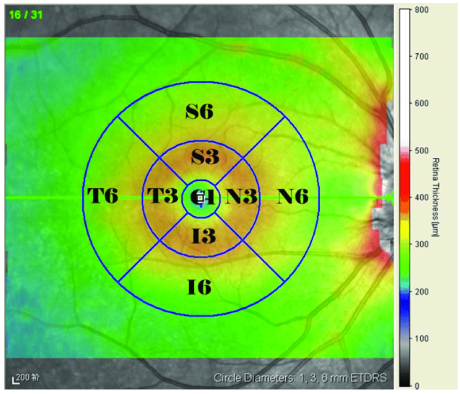


Figure 2.3: Early Treatment Diabetic Retinopathy Study (ETDRS) grids provided by the Spectralis OCT.
Circle diameters 1-, 3-, 6-mm of the central (C1), superior (S3, S6), temporal (T3, T6), inferior (I3, I6) and nasal (N3, N6) sectors. The optical nerve can be partly appreciated the right-hand side of the funduscopic image. Image source: [Li et al., 2017].

The thickness measures used in the final data set were calculated as result of different combinations of regions for each one of the layers, as illustrated in Figure 2.4 (C).

Regions

In total, values of thickness of five layers and the sum of all, corresponding to the total retina thickness, were computed as the average value of both eyes in the different discs and rings: 6-, 3- and 1-mm discs and 1-to 3-mm, 3-to 6-mm rings. All result from the average thickness values of those regions, which are more representative from the anatomical sense [Murueta-Goyena et al., 2019] than those automatically exported by the device.

Layers

It is important to note that two of the seven layers exported automatically were added to their adjacent layer, to add precision to the segmentation, which is often difficult for those cases in which two layers become indistinct. It the case of the GC and IPL (GCIPL) and OPL and ONL (OPONL) layers. Consequently, on the following, thickness values of the mRNFL, GCIPL, INL, OPONL and phot will be mentioned, labelled as mRNFL, GCIPL, INL, OPL-ONL and IS/OS-RPE-BM in Figure 2.4 (B), respectively.

Fig. 2.4 (A) shows the ETDRS grid overlaid on the funduscopic image, to get a better representation of the location of each region on the eye.

Metric protocol

Following Advised Protocol for OCT Study Terminology and Elements (APOSTEL) recommendations [Cruz-Herranz et al., 2016] the measurements of both eyes were averaged

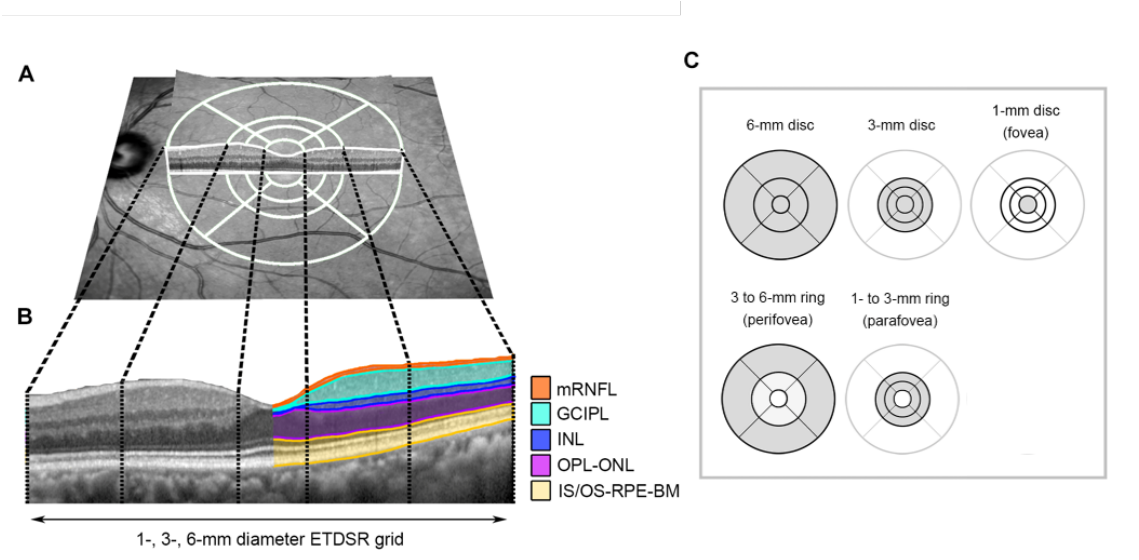


Figure 2.4: Funduscopic image with superimposed foveola-centered 1-,3-,6-mm diameter discs and 1-to 3-mm, 3-to 6-mm rings from the ETDRS grids provided by the Spectralis OCT (A). Layer segmentation of a 6-mm long horizontal B-scan through the foveola (B). Macular regions used for calculating mean layer thicknesses of both eyes (C).
mRNFL, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer, INL, inner nuclear layer, OPLONL, outer plexiform-outer nuclear layer,IS/OS-RPE-BM, inner edge of the junction between photoreceptor inner and outer segments-retinal pigment epithelium. Image source: [Murueta-Goyena et al., 2019].

to account for inter-eye within-patient dependencies unless any pathological condition affected one of the eyes, in which case only the healthy eye was included in the analysis. In this study, this was the case of 3 patients and 5 controls.

3. CHAPTER

Methods - Data analysis

3.1 Preliminary exploratory analyses

3.1.1 Outlier detection

Box plot approach

The first step of the exploratory analysis consisted on the detection and removal of potential outliers. An outlier is a data point that differs significantly from the other data points and should thus be removed. The simplest way to visualize the distribution of a data set is by means of a box plot. This graph displays five statistics: the minimum, first quartile (Q_1), median, third quartile (Q_3) and maximum values of each variable in the data set.

Box plot statistics

A quartile, as its name implies, is a type of quantile that divides all the data points into four equal parts, so that 25 % of the data points lie under the Q_1 value, and 75 % under the Q_3. The inter-quartile range (IQR) is graphically represented in this plot as the length of the box, as each end corresponds to Q_1 and Q_3, respectively. Consequently, the distance in-between each end gives the IQR, which represents in a graphical manner the range of variation of the data.

For its part, the median, shown as a line that falls within the box, corresponds to the middle value of the corresponding ordered data, which represents the value in which the data set is halved in the 50 % highest and lowest values.

Finally, an additional line outside the box starting from each of its ends is drawn to the maximum and minimum values.

Outlier detection rule

A detection of outliers criterion was proposed by [Tukey, 1977], based on the box plot concepts presented in the paragraph above. The proposed method consisted on a simple calculation defined by the IQR, namely the definition of outliers as the points 1.5 times the IQR on the upper and lower boundaries of the boxplot. The data was divided into four boundaries, defined as the two inner and two outer fences by the author. The limiting values of the inner fences were 1.5 times the IQR under/above the first/third quartiles, respectively. A similar logic applied for the outer fences, this case 3 times the IQR above or under $Q1$ and $Q3$.

The points outside the inner fences but inside the outer fences are considered mild outliers, while data outside the outer fences extreme outliers [Dawson, 2011], as shown in Eq. 3.1. The robustness of boxplot test outlier detection method (known as *3σ-rule*) has been analyzed in many studies as, for instance, in [Lehmann, 2013, Andrea, 2013].

$$\begin{aligned} \text{mild outliers} = &\{x_i \leq Q1 - 1.5 \times IQR \quad \vee \quad x_i \geq Q3 + 1.5 \times IQR\} \quad \wedge \\ &\{x_i \geq Q1 - 3 \times IQR \quad \vee \quad x_i \leq Q3 + 3 \times IQR\} \\ \text{extreme outliers} = &\{x_i \leq Q1 - 3 \times IQR \quad \vee \quad x_i \geq Q3 + 3 \times IQR\} \end{aligned} \tag{3.1}$$

for $x_i \in X$ each sample point of the feature X in a d-dimensional feature set.

Only mild outliers were found on this data set by applying the above mentioned rules. Only 8 outliers in controls and 3 in patients were found with the application of the *3σ-rule*, and their values were imputed with the median value. This fact suggests that it consists on data with normal distribution (bell-curve-shaped data).

The box plots for the thickness values in all six layers and regions of the retina is shown in the Appendix section, concretely in Fig. A.1.

3.1.2 Study of the statistical distribution: Normality test

Histogram approach

Some methods applied in the next steps make assumptions on the distribution of the data set, more precisely, a normal distribution is assumed in parametric approaches to make

inferences on the data. The most straight-forward approach to assess whether the data is normally distributed is by looking at its histogram, to get an intuition on how close to a Gaussian distribution it is. The histogram represents the frequency, that is, the number of samples that fall in each non-overlapping consecutive interval that divides the range of values of the data.

The histograms of each variable conditioned to each class group were visualized, to detect any striking non-Gaussian distribution. However, that was not the case for this data set, as shown in Figures A.2 and A.3 for patient and control data, respectively. In general, all distributions resembled to a bell-shaped distribution.

Statistical normality test

In addition to this first approach, a classic statistical test to measure the goodness-of-fit of a distribution's departure from normality has been used: the D'Agostino's K^2 test. Two statistics are used to test normality: skewness (s, Eq. 3.2) and kurtosis (k, Eq. 3.3). The first measures how asymmetric the distribution is: a value of s lower than -1 or greater than 1 indicates a highly skewed distribution, lower values show moderately skewed distributions or symmetric, if s lies between ± 0.5. Kurtosis measures the degree of extremity of deviations (or outliers) and its metric values are analogous to those of skewness.

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{3/2}} \tag{3.2}$$

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{2}} - 3 \tag{3.3}$$

s skewness and k kurtosis of a random variable X with n samples.

The D'Agostino's K^2 normality test combines these to measures, in such a way the Z-statistic takes s and k into account: $Z = s^2 + k^2$). P-values under 0.05 indicate extreme values of Z, suggesting that the random variable X is not normally distributed. This was the case of 5 variables in controls and 1 in the patients, as presented in Table 3.1.

Similarly, the results of the normality test for the remaining variables are shown in Table A.2 (control data) and A.1 (patient data).

Layer	Region	Z-statistic	P-value
Patients			
GCIPL	1 mm	6.779	0.034
Controls			
GCIPL	1 mm	9.143	0.01
	1 mm	6.51	0.039
INL	1 3 mm	8.961	0.011
	3 mm	9.645	0.008
mRNFL	3 6 mm	6.307	0.043

Table 3.1: Variables without a normal distribution, according to D'Agostino and Pearson's omnibus test of normality [D'Agostino, 1971].
Table values presented by layer and region showing $Z - statistic$ and associated $p - value$ of the normality test.

3.1.3 Study of feature correlations

Correlations and redundancy implications

The final consideration in this preliminary exploratory analysis of the data was to find the degree of correlation between features, as high values of correlated features may indicate the existence of redundant features, which may deteriorate the predictive performance of the learning algorithm.

Features that are highly correlated may result redundant, in the case that both have the same or really similar number of positive and negative examples [Appice et al., 2004].

Mitigation approaches

One way to deal with this unnecessary added level of complexity to the classification problem could be the elimination of those features (**feature reduction**), as several algorithms based in pairwise comparison of the features (*REDUCE* algorithm [Lavrač et al., 1999, Appice et al., 2004]) or on detection of such features using greedy accuracy-based heuristics (GREEDY3 and GROVE algorithms) or decision trees (*FRINGE* algorithm) do.

Nonetheless, the most widely investigated approach is the one which aims to find the most *relevant* (**feature selection**) but at the same time non-redundant set of features (i.e. multivariate feature selection). This is the main topic of Section 3.2. Some of the methods for feature selection take into account the feature dependencies, but also prioritize the correlation with the class when selecting the most relevant predictors.

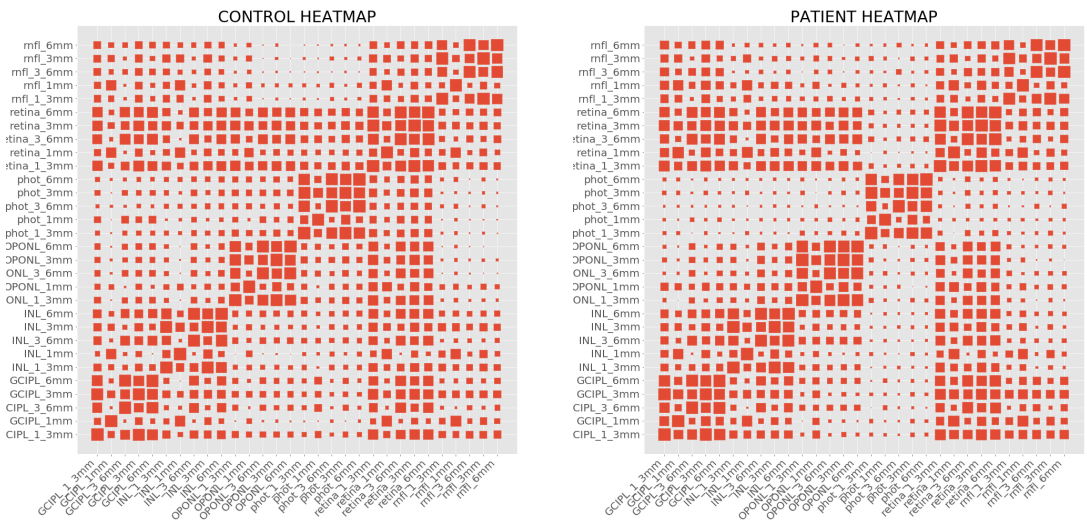
Pearson correlation

The most common approach to measure the linear correlation between two variables is the Pearson correlation coefficient (ρ_{X_1,X_2}). Features that are highly positive or negative correlated have closer values to ± 1, respectively. Totally uncorrelated features in turn would give correlation values closer to zero. Eq. 3.4 shows the formula to compute pairwise correlation values for two predictors $\{X_1, X_2\}$.

$$\rho_{X_1,X_2} = \frac{\mathrm{E}\left[(X_1 - \mu_1)(X_2 - \mu_2)\right]}{\sigma_1 \sigma_2} \tag{3.4}$$

where $\mu_i = E[X_i]$ is the expected mean value of the X_i variable and σ_i the standard deviation of X_i.

The feature-feature correlations of the data set were analyzed with the Pearson correlation coefficient. In Figure 3.1, the features with higher correlation values are clustered together. The correlation values are represented in this figure as the square size. As expected, the variables from the same layer but different region have a higher correlation than with those from other layers and appear clustered together.



(a) Control data **(b)** Patient data

Figure 3.1: Correlation plot based on Pearson correlation values of control (a) and patient (b) data. rnfl, retinal nerve fiber layer; phot, photoreceptor layer; OPONL, outer plexiform-outer nuclear layer; INL, inner nuclear layer; GCIPL, ganglion cell-inner plexiform layer; {6-,3-,1-mm}, 6-,3-,1-mm discs; {1_3mm, 3_6mm}, 1-to 3-mm and 3-to 6-mm rings.

A correlation plot similar to the one in Figure 3.1 but with explicit Pearson correlation

values is shown the Appendix, in Figure A.4.

3.2 Feature selection

Introduction to the problem

The general definition of feature selection (FS) in machine learning (ML), or variable selection, is the process of selecting a subset of the most relevant features prior to the model construction.

The objective of such a selection is three-fold: improving the prediction performance of the ML model, providing faster and more cost-effective predictors, and grating a better understanding of the underlying process that generated the data [Guyon, 2003].

A formal definition to the above-stated problem: let D be the feature set with a large number of features $X_1, X_2, ..., X_d$ where d is the number of features. FS is defined as the process of selecting the k most discriminatory features out of the d-dimensional feature space, such that $d \geq k$ [Haindl, 2006].

Areas of application

FS has become the focus of many disciplines which rely on large datasets for training ML models, including biomarker discovery or microarray gene expression data classification problems in bioinformatics, image processing or text mining with high-dimensional feature spaces.

Motivation of its application in this study

In this project, the application of these techniques was fundamental to avoid the problem of model overfitting. Overfitting may occur due to fitting a model which is too complex for the data used in training, which will then not generalize to unseen data and will, consequently, yield poor validation results. The risk of overfitting is greater with small sample sizes. In addition to this reason, a higher number of features exacerbates the problem of overfitting, as the probability of finding a combination of features that discriminates the class variable purely by *chance* is higher.

The number of features in this dataset was not extremely high, compared with the ones used in the above mentioned disciplines. Nevertheless, it is relevant given the limited sample size of the data. The work in [Jovic et al., 2015], reviewed several FS techniques

applied to different disciplines and proved that feature selection is essential in any process learning from data.

Methods

Feature selection can be used in supervised and unsupervised learning tasks. When it comes to labelled instances, many approaches exist and are useful. The methods here used are presented following the taxonomy of the feature selection techniques review within the domain of bioinformatics introduced by the authors in [Saeys et al., 2007]. Depending on how the features are combined to construct the classification model, these methods are organized into three categories: filter methods, wrapper methods and embedded methods. The first two are used in this work.

3.2.1 Filter methods

Filter methods select features based on measures that are independent of the employed data modeling algorithm. That is to say, it consists of techniques that rely only on the properties inherent to the data.

There are different classes of filter methods, not only dependent on the task (classification, regression or clustering), but also on how the best features are found. The latter involves univariate or multivariate filters. Univariate filters evaluate and usually rank every single feature independently, while multivariate filters evaluate an entire feature subset.

Univariate Filters

The principle is as follows. Based on the score of each feature independently obtained from a metric of interest, a ranking of the best features is built. From it, the ones with lower scores may be removed. This is the principle of the *K-Best Feature Selection*.

In this project, the following metrics have been applied to the data:

1. Unsupervised filters:

    - Variance-threshold filter(σ^2). This method is built upon the variance of each variable, by removing the ones under a certain threshold value. This way, the ”quasi-constant features", i.e. features with many similar values in all their instances, are considered irrelevant prior to constructing the model.

$$\sigma^2 = E[(X-\mu)^2] \tag{3.5}$$

The variance of a random variable X is the expected value of the squared deviation from the mean of X, $\mu = E[X]$.

2. Supervised filters:

   - Parametric approaches

     – Welch's T-test (t). The score is based on a two-sided T-test to determine if the (expected) average values of two independently sampled populations are different under the null hypothesis. The test is insensitive to the equality of variances, as opposite to the original definition of the Student's t-test, which is only robust in the presence of unequal variances with comparable sample sizes [Markowski and Markowski, 1990].

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{3.6}$$

     where μ_j is the mean value of each distribution, σ_j the standard deviation and N_j the sample size of each distribution, j $\in \{1,2\}$.

     – ANOVA Analysis of Variance ($F - ratio$). One-way ANOVA uses the F-distribution to test if two or more groups are significantly different by comparing variances, similarly to the t-test which compares the *difference* between the mean of two populations. ANOVA tests for the *variation* between the means of two (or more) groups and within the mean of the groups.

     The advantage of this method appears only when comparing more than two groups simultaneously, as running multiple t-tests increases the Type I error probability [Kao and Green, 2008]. While the t-test would require a lot of pair comparisons, the method used in ANOVA finds an overall difference between three or more means by using a single test that compares all the groups simultaneously, the *f-statistic*. Nonetheless, only two groups are considered in this study, so both are equivalent. In fact, the F-ratio in such case equals the square of the t-statistic.

     In Eq. 3.7 the F-ratio is calculated for a one-way ANOVA based on the sum of the squared deviations of observations from the mean: *Sum of Squares* (*SS*), such that for a sample i-th from the j-th group (x_{ij}), its $SS = \sum \left(x_{ij} - \mu\right)^2$.

$$F - ratio = \frac{MS_B}{MS_W} \tag{3.7}$$

It is calculated from the *Mean Square* (*MS*) of the *SS within* (MS_W) and *between* (MS_B) groups. The *MS* refers to the average squared deviation of observations from the grand mean, i.e. the mean of the entire population, and it is obtained by dividing *SS* by the degree of freedom of the population (df). For instance, $MS_B = SS_B/df_B$. The df of a population of size n is $n-1$, because those are the number of observations required to obtain the grand mean.

- Non-parametric approaches:
  - Mann-Whitney U test or Wilcoxon-Mann-Whitney test. The features are ranked according to the statistic value test on two continuous and ordinal distributions with unpaired samples.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{3.8}$$

The null hypothesis tests the probability of a randomly drawn observation (X_1) from one group to be larger than a randomly drawn observation from the other (X_2) [Hodges and Lehmann, 1963]. A numeric rank is assigned to all the observations of one set, beginning with one for the smallest value and a midpoint value to handle ties. The value of the U-statistic (U_1) is shown in Eq. 3.8, where n_1 is the sample size and R_1 the sum of the ranks in sample one. The same would apply for the second set (U_2).

  - RelieF algorithm (W). This algorithm is based on the capability of the attributes to differ between instances that are close (low distance between their values), assuming no conditional independency of the attributes with the class [Urbanowicz et al., 2018].
Given a d-dimensional feature space, the algorithm assigns a weight (W for feature X_k; $W[X_k]$) to the attributes by taking into account the distance with the k closest instances from the same class (*near Hit,* H_j) and the closest from the other class (*near Miss,* M_j), as shown in Eq. 3.9.

$$W[X_k] = W[X_k] - \frac{\text{diff}(X_k, s_i, \text{H}_j)}{(m*k)} + \frac{\text{diff}(X_k, s_i, \text{M}_j)}{(m*k)} \tag{3.9}$$

$$k \in \{1,2,...,d\}, \quad i \in \{1,2,...,m\}, \quad j \in \{1,2,...,n\}, \forall i \neq j.$$

For continuous variables the feature score vector W is updated based on the feature value differences observed between m randomly selected instances (s_i) out of the n total number of samples ($m \leq n$) and their k neighboring instances, which are either from the same class (*hits* H_j) or the other (*misses* M_j).
Features (X_k) that have different value between the randomly selected instance s_i and their miss neighbors M_j are considered informative, while a different value with their hit neighbors H_j gives evidence to the contrary. The equation 3.10 gives a numeric score which accounts with the aforementioned logic.

$$\text{diff}(X_k, x_i, x_j) = \frac{\left|\text{value}\,(X_k, x_i) - \text{value}\,\left(X_k, x_j\right)\right|}{\max(X_k) - \min(X_k)} \tag{3.10}$$

for a feature (X_k) in the d-dimensional feature space, the distance between the instance s_i value in that feature (x_i) with either its hit or miss neighbor (x_j).
In short, the weight values for each feature X are updated with the distance values ($\text{diff}(X_k, x_i, x_j)$), computed by means of the Euclidean distance of the samples of a randomly selected feature instance with its closest hit and miss neighbors.

Finally, this last two presented methods required the **discretization** of the data. These methods work on categorical data, so each feature X_i was discretized into three equal-size buckets following the quantile criteria, e.g terciles. By using this method based on the *frequency*, it assures that no empty bins, but bins with same or similar number of samples, are formed.

- Chi-square test (X^2). The chi-squared distribution is used to test whether a statistically significant difference between the expected and observed frequencies of a categorical variable exists [Cochran, 1952]. That is, the discrepancy between the observed (O_i) and theoretical frequencies (E_i) 3.11, known as *goodness of fit*.
  If the null hypothesis is true, the observations follow a chi-squared distribution in their limit distribution and their frequencies do not differ significantly.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.11}$$

The expected frequency O_i is calculated upon the probability of that class to fall into that class (p_i), such that for n samples, $O_i = n * p_i$.

– Mutual Information (MI). In information theory, the mutual information between two random variables, e.g. the features X and the class variable Y, quantifies the amount of information obtained about one random variable through the other.
The amount of information is given by the entropy, which is a measure of the *uncertainty* of a random variable. The more uncertain its value is, the more information it gives [Shannon, 1948].
Although continuous data is also supported to compute the entropy in *scikit-learn* package, by using a metric based on the distance of k nearest neighbors of a sample point s_i in Y to calculate the MI with the number of neighbors in X that fall in that area and are from that same class in respect to the number of neighbors from all the possible classes within that area [Ross, 2014].
However, as the data was previously discretized analogous to the chi-squared method, the MI was computed from the contingency table, as shown in Eq. 3.12.

$$\begin{aligned} \mathrm{MI}(X;Y) &= \sum_{i=1}^{|X|}\sum_{c=1}^{|Y|} P(X_i,Y_c)\log\frac{P(X_i,Y_c)}{P(X_i)P(Y_c)} \\ &= \sum_{i=1}^{|X|}\sum_{c=1}^{|Y|} \frac{|X_i\cap Y_c|}{n}\log\frac{n|X_i\cap Y_c|}{|X_i||Y_c|} \end{aligned} \tag{3.12}$$

where $|X|$ is the number of levels of the variable X and $|X_i|$ the number of samples that fall into each level. Same applies for the Y variable.
As a result, the MI measures the degree of dependency between the variables with the class, so the rank of features is built based on a higher value of MI.

Multivariate Filters

Introduction

Multivariate filter methods seek to integrate also dependencies between the features into

the feature selection process and search for, instead of a ranking of features, the best feature **subset**.

To generate it, several strategies have been proposed in the literature known as search methods, due to the fact that the search for the best subsets yields a NP-hard combinatorial problem, so conventional numerical methods are unfeasable. *Search methods* are generally divided into three categories: exponential algorithms, sequential algorithms and randomized algorithms. Additionally, a *metric score* needs to be defined, in order to evaluate each feature combination.

As a final remark, these algorithms return thus a suboptimal solution, which allows to give an approximation of the best feature subset without checking the goodness of fit of all the possible subsets.

Method: Correlation-based Feature Selection (Merit$_S$)

A selection of this type was conducted following the method proposed already by the authors in [Hall, 1999]: the Correlation-based Feature Selection (CFS). The algorithm is based on taking into account the usefulness of individual features for predicting the class label (as univariate filters do), but also the level of intercorrelation among them.

Metric score

To evaluate the worth of the subset (S), CFS uses a correlation based heuristic called *merit*:

$$\text{Merit}_S = \frac{k\overline{r_{XY}}}{\sqrt{k+k(k-1)r_{XX}}} \tag{3.13}$$

The Merit$_S$ of a feature subset S of k features, is the metric shown in Eq. 3.13, where $\overline{r_{XY}}$ is the mean feature-class correlation and $\overline{r_{XX}}$ the average feature inter-correlation.

Feature correlations are estimated based on the information theory via the Information Gain (IG), which measures the amount of information by which the entropy of Y decreases provided X, as the method that the authors in [Zhao and Morstatter, 2010] proposed. More specifically, CFS calculates the feature-class and feature-feature correlations using symmetrical uncertainty (SU) (Eq. 3.14), to mitigate the bias in favor of features with more values that occurs if only using the information gain.

$$\text{SU}(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right], \tag{3.14}$$

where $IG(X|Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$ is the information gain of a feature

X given the class labels Y and $H(X), H(Y)$ are the *entropy* of the variables [Shannon, 1948].

Search method and stopping criterion

Finally, CFS explores the search space using the Best First search with a stopping criterion of five consecutive fully expanded non-improving subsets. The idea is to introduce the maximum relevant feature while avoiding the re-introduction of redundancy by recalculating the SU of each feature.

3.2.2 Wrapper methods

Method and metric score

Wrapper methods address the problem of feature selection as a search problem, in which the output of the selection is not a ranking of features, but a feature subset. Subsets of different combinations of features are prepared, evaluated and compared to other combinations.

The *metric score* used to compare them is given by a metric of performance given by the predictive model. Despite the many metrics available to measure classification performance [Hand, 2012], this work focused on evaluating the performance based on a classification accuracy criteria, i.e. how well the classifier assigned objects to their correct classes.

Limitations

Hence, in contrast to filter methods, the selection is not independent from the data modeling algorithm, which has some negative implications. The most critical: being computationally more costly and having a higher risk of overfitting. However, it is not only because it accounts feature dependencies as multivariate filters do, but also because it considers the interaction between the feature subset and the model selection, why this method adds so much value to solving the problem of feature selection [Saeys et al., 2007].

Search methods

Regarding the *search methods* used to find the feature subsets, this work focused primarily in sequential algorithms, but not only. It also considered the solutions provided by an exhaustive search to find feature subsets, whose complexity grows exponentially with the number of features k considered in the subsets. Due to its complexity, which makes it unfeasable in terms of computational time, the dimensionality of the feature subsets were constrained to only two or three features (bounded exhaustive method).

For the case of sequential algorithms, greedy search methods seem to be particularly computationally advantageous and robust against overfitting. The sequential forward and backward greedy search methods are considered in this work.

The following algorithms are implemented in the *mlextend* Package [Raschka, 2018].

1. Sequential

   - Greedy forward feature selection (SFS). The algorithm starts from an empty feature subset and adds one new feature at each iteration, the one that yields the highest performance score upon addition. At each iteration k_i, one new feature is included in the subset and the performance score is saved, regardless of it being higher or lower than in the previous iteration $k_i - 1$.

     It is important to note here, that the maximum number of features to include in the search process is fixed by the user, so the number of features that produces the best subset can selected out of all the classification scores at the end of the search process.

   - Greedy backward feature selection (SBS). The process here is the reverse as in the one above. The algorithm starts with a subset including all features, and removes at each iteration the feature that yields the best performance improvement upon its removal.

     Similarly to the SFS selector, the best feature subset, if any, is selected at the end based on the highest performance score achieved during the whole search process.

2. Bounded exhaustive

   - Exhaustive feature selection at K dimensionality (EFS). Exhaustive search methods implicate the evaluation of all the possible combinations of features. The complexity of this problem grows exponentially, such that the complexty of an exhaustive search with n features:

$$O_{k_i}(n) = \frac{n!}{k_i!(n-k_i)!} \tag{3.15}$$

     for each feature subset with dimension k_i, $i \in \{1,2,...,k\}$.

Due to the number of features in this dataset, the approach is unfeasable. Nevertheless, the problem was solved for feature subsets with dimensions restricted to two ($O_2(30) = 435$) and three features ($O_3(30) = 4060$).

3.2.3 Limitations

Bias in ML models

Feature selection is an essential part of the machine learning process, as important as any other step, like model selection. Therefore, some considerations about the algorithms used for this task should be mentioned. As it is not commonly the case that the problem implicates a low number of features for which classical solving methods may apply, a metaheuristic approach is generally the best possible option to solve this NP-hard problem.

Filter methods

On the one hand, different methods give different type of results. Although the properties of the features give relevant information that should be accounted for in order to make inferences on the data, the dependencies between them play also a substantial role, so the search of a feature **subset** appears as the main goal. Multivariate filters and wrapper methods give this kind of solution to the problem. In addition to that, multivariate filters produce a subset of features that are not tuned to a specific model, so even if the result is more general, these feature sets give lower prediction performance than wrappers do [Zhang et al., 2013].

Wrapper methods

On the other hand, wrappers are much slower than filter methods, so they require fast modelling algorithms such as Naïve Bayes or Support-Vector Machines (SVM). Additionally, these methods could be biased towards the modeling algorithm on which they were evaluated. To obtain a reliable estimate of the generalization error, an independent validation sample is required. Feature selection should be performed within the model selection process, as otherwise it would introduce bias and result in an overfitting of the training data [Kuncheva and Rodríguez, 2018]. The result would be seen as an outperformance compared to the other ML models, but only because the same training data was also used to choose the best subset based on that specific model, giving unreliable results in validation.

Stability of selected features (I_C)

Another limitation of the method is that we do not have the real value of accuracy, but only an approximation on the training data to add or remove features from the subset. The size of the data set limits the number of possible data shuffling turns, so it is not possible to obtain an estimate of the accuracy with unequivocal variance.

Factors of stability

The result is a degree of variability between the subsets that calls for attention. The consistency between pairs of subsets can be measured in terms of the length of the intersection of the subsets (*monotonicity*), by using a stability index with a value bounded by constants that are independent of the total number of features (d) and the cardinality (k) of the subsets (*limits*) and that is constant for independently drawn subsets (*correction for chance*) [Saeys et al., 2007].

Index metric

Considering the cardinality of the interesection of the subsets as $r = |S_i \cap S_j| \leq k$ for subsets S of k features drawn from a d-dimensional feature space, such an index is shown in Eq. 3.16 (Equation (2) in [Saeys et al., 2007]).

$$I_C(S_i, S_j) = \frac{r - \frac{k^2}{d}}{k - \frac{k^2}{d}} = \frac{rd - k^2}{k(d-k)} \tag{3.16}$$

$\{i, j\} \in \{1, 2, ..., m\} \wedge \forall i \neq j$, m is the total number of subsets S considered.

In Eq. 3.17 this index is introduced in a score to measure stability between a set of m subsets of k features ($S = S_1, S_2, ..., S_m$). It consists indeed on the average of all the pairwise consistency indices between two feature subsets of k features selected out of the d total number of features.

$$\mathcal{I}_S(\mathcal{S}(k)) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} I_C\left(S_i(k), S_j(k)\right) \tag{3.17}$$

where $I_C(S_i, S_j)$ the consistency index between the two subsets from the m subsets S of k features, $\{i, j\} \in \{1, 2, ..., m\} \wedge \forall i \neq j$.

Applications

This metric is really interesting to choose the final sequence of features. Then, if the stability is high ($I_S \geq 0.5$), a rank of features ($S*_{rank}$) based on performance as the score metric appears to be better than choosing the subset with the minimum error ($S*_{minE}$) found during training. One way proposed by [Kuncheva, 2017] to build $S*_{rank}$ consists on assigning weights to the features based on their position in the different subsets (so that the first feature is the one in the first position), ordering them in ascending order according to the sum of their values in all subsets and validating the rank on the testing set.

This approach may be of interest, on the one hand, to find consistency between the feature subsets obtained in different training-folds of the validation procedure. On the other hand, to find consistency between the different subsets extracted from different feature selection methods.

Stability concepts are applied in this study to gain deeper insight of the feature subsets when applying univariate filters for feature selection. In the results section, the maximum subset intersect between all folds (S_{iMax}) and its cadinality (r), the number of features that first yield best performance on each cross-validation fold (k_{best}) together with the total number of folds with that k_{best} number (s) and the selected final subset (S_{best}) are reported. Note that the $\sum s_i = 10$, i={1,...,z}, where z is the total possible values that k_{best} can take in each 10-fold cross-validation.

To select this one final subset S_{best}, two options apply:

- S^*_{rank} **is reported**. There is enough consistency in the number of features with the highest performance between all the folds, i.e. coherence in k_{best}.

  Consistency is considered only if two conditions are fulfilled: (a) there exists a maximum subset intersect S_{iMax} sufficiently large ($r \geq k_{best}$) and (b) there exists enough accordance between the folds, i.e. coherence in s (k_{best} yields $S^*_{rank} \iff s_{k_{best}} > 6$).

  Then, and only then, a subset S^*_{rank} that is generated from the first k_{best} features of S_{iMax}, is the one reported.

- S^*_{minE} **is reported**. The feature subsets generated on the cross-validation are not consistent.

  Such cases only occur in one of the following alternatives: (a) there are more than 2 possible number of features with best performance, i.e. more than two possible values of k_{best}; (b) k_{best} can take only two possible values but there exists a tie

between them ($s = 5$), (c) the coherence between folds is lower than k_{best} number of features with best performance ($r < k_{best}$).

If stability is low, selecting the best individual run is more useful [Kuncheva, 2017]. Therefore, the subset(s) from the fold with highest performance are reported, that is, S^*_{minE}.

These tables allow to report features that play a role in the different filter feature selection methods, the level of accordance of each filter in the cross-validation and a procedure to select the best feature subsets. However, dealing with tides when selecting the optimal feature number in each fold appears as the main limitation. Selecting the first number with best performance was the strategy used in this work. Hence, the complementary figures of mean score values and standard deviation are fundamental in order to draw conclusions of the method and its results.

3.3 Supervised classification

Introduction

The problem addressed in this work is a binary classification problem, in which the goal is to find the features that are more discriminant to discern between controls and patients and learn a supervised learning model to accurately discriminate the problem phenotype.

The feature selection methods presented in Section 3.2 are also part of the machine learning classification task. It was only due to their extension and protagonism in the project that these are reported in an independent section. Accordingly, this section aims to present the classifiers used in this work's experiments and a short insight of their main principles, as well as their integration with the feature selection step and the validation of performance of the resulting system.

3.3.1 Dealing with imbalanced data

Before the aforementioned is presented, a brief introduction and motivation to the use of oversampling techniques in this work. Accuracy is used as a standard performance evaluation metric in many classification problems. However, this metric gives same weight to all misclassification errors, whereas in this work, the false positive rate has more clinical

relevance than the contrary, i.e. incorrectly assigning a healthy subject to the patient group (false negative rate).

Another aspect to take into account in feature selection techniques is that the selection of the most relevant variables are based on the given score metric. For instance, information gain does not give more relevance to a certain class, but it is computed for each sample of each class. As the authors in [Chawla et al., 2002] claimed, in an imbalanced data set where the sample of the minority class is less than half of the one from the majority class, most of the features will be associated with the negative class.

To tackle this bias problem, two main approaches have been proposed in the literature. On the one hand, undersampling techniques approach the problem by reducing/deleting instances from the majority class. On the other hand, oversampling techniques duplicate examples from the minority class until a more balanced distribution is reached.

This work used the second mentioned type of techniques, as undersampling may reduce the classification ability due to the elimination of data rich on information of the majority class [He and Ma, 2013]. Precisely, the SMOTE oversampling technique presented by authors in [Chawla et al., 2002] was selected to create a balanced dataset. Whereas random oversampling may increase the likelihood of overfitting and decrease the classifier performance [Fernández et al., 2018] – since it makes exact copies of the minority class – SMOTE creates synthetic training data by interpolation of several minority class instances that lie together. To do so, the algorithm generates new samples for N randomly selected instances of the minority class by interpolation, i.e. computing first the distance (dif) with one of its k neighbors selected at random and secondly adding a gap selected between 0 and 1 times the distance to the value of that N_i selected instance to create the corresponding synthetic sample ($synthetic_{N_i}$). This interpolation mechanism is shown in Eq. 3.18:

$$\begin{aligned} dif = &\quad N_i - random(k) \\ gap = &\quad random(0,1) \\ synthetic_{N_i} = &\quad N_i + gap * dif \end{aligned} \tag{3.18}$$

for each one of the i randomly selected samples of the minority class (N) and its corresponding k neighbor instances, to generate i synthetic new samples of the minority class in the training set ($synthetic_{N_i}$).

In this work, the number of neighbors was set as default ($k = 5$) and to equalized the number of samples was the resampling strategy of the SMOTE technique.

3.3.2 Learning algorithms

Four classification algorithms were considered for supervised classification in order to compare the results of the pipeline presented in Sections 3.1 and 3.2 when different models are used. In this work, the values of the parameters of the classification algorithms were fixed by default, as the work did not lay focus on the hyperparameter optimization or tuning of the model, but on the properties of the data.

The cornerstone of the algorithms used for this work are hereby presented:

- K-Nearest Neighbors Classifier (KNN). The class membership is assigned to the most common class among its k nearest neighbors, that is, the ones with lowest Euclidean distance, as it was the case of continuous variables. The features of this work had numeric values, so the Euclidean distance was selected although other metrics may apply for this method.

  A limitation of the KNN method is that it is instance-based learning or non-generalizing learning. This means that it does not construct a model, but works only on the stored training data. To overcome the effect of ”majority voting" that occurs in skewed situations, different approaches have been proposed to give weights to increase the importance of the closest neighbors [Dudani, 1976] and reduce, additionally, the effect of outliers, which affect negatively to Dudani’s distance-weighted algorithm in the case where an outlier is closer to the test element than the training data of that same class, what adds bias to the prediction [Gou et al., 2012].

  In this work, a uniformly weighted KNN classifier - $k = 5$ neighbors - was selected, to prevent giving higher weights to neighbors which belong to highly correlated features and have, therefore, similar distance values with the element to add a class membership. An uneven number of neighbors is preferred, to avoid tides occurring as the neighbors are equally weighted.

- Naïve Bayes Classifier (NB). The classification relies on the maximum-likehood criterion to construct a Bayes probability model as the classifier (Eq. 3.19) and the maximum a posteriori rule (*MAP*) as the decision rule (Eq. 3.20), which selects the most probable hypothesis when assigning a class label to a given instance.

This classifier is naïve, because it assumes that all features are mutually independent given the class variable-phenotype when building the joint probability model. The chain rule is used to determine the conditional distribution over the class variable during the model construction. The estimation of the probability of an instance to be in a class is based on its the relative frequency estimated in the training set, which makes the classifier computationally fast compared to more sophisticated methods but its probability outputs to have low reliability [Zhang, 2004].

The Naive Bayes classifier rule corresponds to the Bayes probability model, and is shown in Eq. 3.19. In this equation, only the numerator is of interest (joint probability model), as the denominator does not depend on the class and is constant.

$$p\left(Y|X_1,\ldots,X_d\right) = \frac{p(Y)p\left(X_1,\ldots,X_d|Y\right)}{p\left(X_1,\ldots,X_d\right)} \tag{3.19}$$

Finally, the decision rule of the model (*MAP*) is presented in Eq. 3.20:

$$\text{classify}\left(x_1,\ldots,x_d\right) = \underset{c}{\text{argmax}} p(Y=y)\prod_{i=1}^{d} p\left(X_i=x_i|Y=y\right) \tag{3.20}$$

- Decision Tree Classifier (Decission Tree). The predictive model, called classification tree, draws conclusions from the feature observations of a target variable, represented in the branches, to the class label, represented in the leaves. The branches are the conjunction of features that lead to those class labels. To construct this structure, a set of splitting rules of the whole set based on the features is used in a recursive manner (recursive partitioning). The data set is split until the subset has the same target variable for all its instances, or when splitting no longer adds value to the precitions [Shalev-Shwartz, 2014].

  The decision rules are inferred from the data, so the result is an observable white box model. The deeper the tree, the more complex decisions and the fitter the model is. To measure the quality of a split, the Gini impurity was used, although the entropy for the information gain is also supported in *scikit-learn* library. Gini impurity measures the level of miss-classification of a randomly chosen instance to be incor-

rectly labeled according to the distribution of labels in the subset.

$$I_G(p) = \sum_{i=1}^{X} p_i \sum_{k \neq i} p_k; \sum_{k \neq i} p_k = 1 - p_i \tag{3.21}$$

where I_G is the Gini impurity for a sate of items X, p_i the fraction of items labeled with class i and p_k the probability of a mistake in categorizing that item.

- Support-Vector Machine Classifier (SVM). The SVM model represents the n instances of each feature X as points in a higher-dimensional space where a segregation by class can be easier. To separate them as wide as possible, the SVM defines a maximum-margin hyperplane, that servers as the classification decision boundary.

  The search of a classification boundary by adding new dimensions can be computationally terribly expensive, so instead of working with feature vectors, SVM uses the dot products between them, to define a linear or nonlinear kernel function that separates them. In the resulting high dimensional space, the dot products of pairs of input data can be computed easily in terms of the variables in the original space [Press, 2007].

  The features in the high dimensional space x are mapped to the points in the original space x_i using combinations with α_i parameters and the kernel function. The function of the hyperplane is defined by the set of points whose dot product with a vector in that space is minimal. Such a transformation is shown in Eq. 3.22 for the case of features that are linearly separable-the decision boundary is lineal.

$$\sum_i \alpha_i k(x_i, x) = \text{ constant} \tag{3.22}$$

  The high-dimensional hyperplane constitutes the decision boundary, such that new features are assigned class labels depending on the region where they fall into.

3.3.3 Validation pipeline

Introduction of validation methods

Validation methods consist on the generalization of the prediction, i.e. out-of-sample performance prediction. But they also cover the model selection: to determine the number of significant variables, to guide and halt the search of competitive variable subsets, tune

or set by default the hyperparameters of the model and, only then, evaluate the final performance of the system. For the last purpose, a set of the data must be reserved as an independent set test and the rest of the tasks be carried on the data for training.

Validation method

In this work, a 10-fold cross validation was used to provide an estimate of the model performance, particularly the ones described in Section 3.3.2. A cross-validation is a re-sampling procedure that consists on randomly dividing the dataset in k groups or folds and to treat each one of them one time as the validation set. The validation set is the one used to evaluate the performance of the classifier, once it has been trained with the other $k-1$ folds.

Cross-validation parameters

The choice of k=10 in the cross-validation was based on a bias-variance trade-off. It has been proven empirically that test error rate estimates when using 10-fold cross-validation suffer neither from excessively high bias nor high variance [Gareth James, 2014]. Bias is higher the more difference exists between the training set and the re-sampling subsets. Higher k values yield smaller test sets, but the training set resembles more the whole data set, so that bias diminishes. However, validation results across smaller test sets also varies more in those cases.

The goal is choosing a k such that the training and testing sets are statistically representative of the broader dataset.

Report of validation result

The results are summarized as the mean $\pm$ standard deviation accuracy scores (accuracy, Eq. 3.23) of each fold of the cross-validation. Validation based on a cross-validation grants more robust results and less biased and optimistic than the ones provided by a simple train/test split [Cawley and Talbot, 2010]. Therefore, for each classification system reported, the performance is assessed through an average of the results in each fold.

$$\text{accuracy} = \frac{TP+TN}{n} \tag{3.23}$$

where TP and TN represent the number of correctly classified instances, i.e. true positives and true negatives, respectively. n represents the all the classified instances, from the positive and negative classes.

Classification score metrics

The problem of misclassification with unbalanced data is that not all classification errors have the same relevance given a certain problem. For instance, in the case of clinical data where Parkinson's disease instances are regarded as the positive class, the importance of incorrectly assigning a patient to the healthy group – called a false negative – is much more harmful, more expensive, than conversely, i.e. incorrectly assigning a control to the patient group (false positives).

For this reason, two more statistics from the confusion matrix, the recall and precision scores, were computed in the wrapper feature selection. Recall has been widely used in information retrieval, as it represents the ability of the classifier to correctly assign the labels of all the instances from the positive class. Finally, precision penalises the false positive rate to test the reliability of the detection system to correctly assign the positive class labels.

$$\begin{aligned} \text{recall} &= \frac{TP}{TP+FN} \\ \text{precision} &= \frac{TP}{TP+FP} \end{aligned} \tag{3.24}$$

where TP, FP, FN refer to the true positive, false positive and false negative rate respectively.

Recall is specially valuable when one class is more relevant and there are few samples of that class. Similarly, precision is very useful if a false positive classification is particularly costly. However, these metrics tend to neglect the evaluation of the prediction ability on the other class. To account for the bias towards the majority class or to the class considered in the recall and precision metrics, [Santafe et al., 2015] described a combination of these metrics to obtain a trade-off between the classification ability on both classes. These metrics are referred as *balanced scores* and consist on the operations shown in Eq. 3.25.

The arithmetic and geometric mean were used in this work to combine the recall scores estimated from the majority and the minority classes. In addition to the recall estimated from the majority and minority classes, the former metrics were applied to evaluate the classification ability of the original dataset and the one with equal proportion in training generated by the SMOTE technique introduced in Section 3.3.1.

$$\begin{aligned} recall\_A\_mean &= \frac{recall_{maj} + recall_{min}}{2} \\ recall\_G\_mean &= \sqrt{recall_{maj} * recall_{min}} \end{aligned} \tag{3.25}$$

where *recall_A_mean* and *recall_G_mean* are the arithmetic and geometric means, respectively, of the recall score calculated on the majority ($recall_{maj}$) and minority ($recall_{min}$) classes.

4. CHAPTER

Results

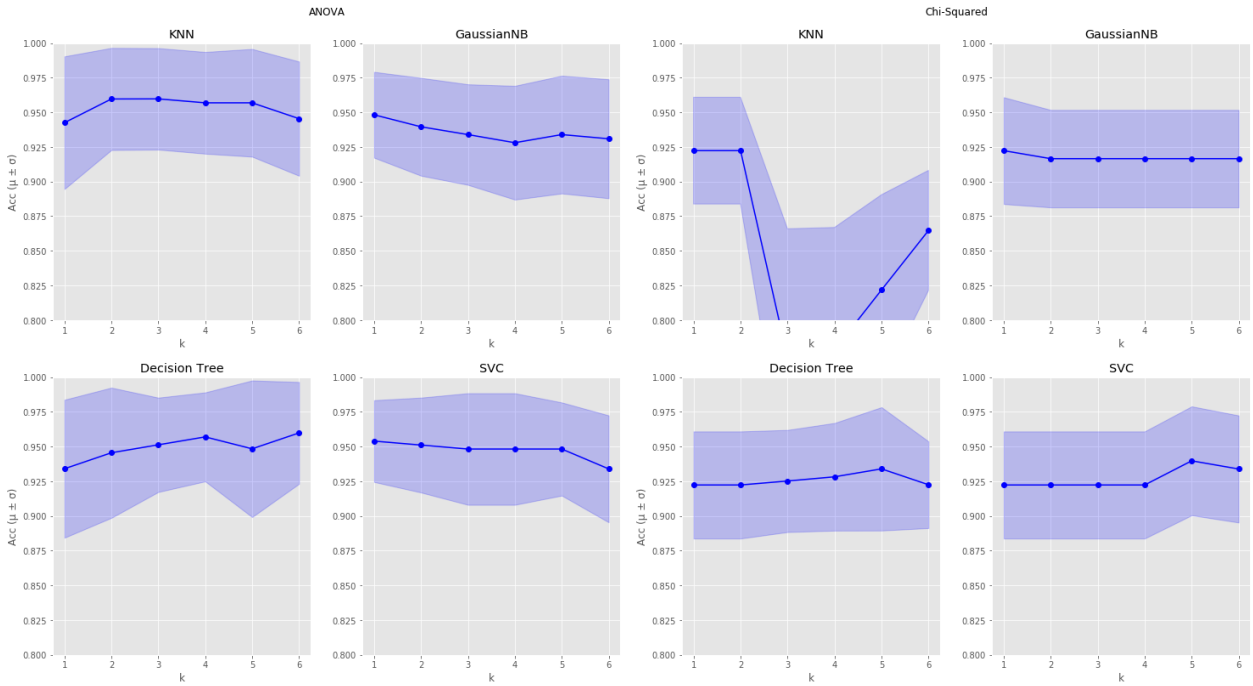
4.1 Feature selection

The aim of this section is to evaluate the information, relevancy and redundancy of the variables in the data set using the methods introduced in Section 3.2 for feature selection.

The results of the 10-fold cross-validation consist of several performance scores reported with its mean (μ) and standard deviation (σ) values. On each experiment, four different classifiers were used.

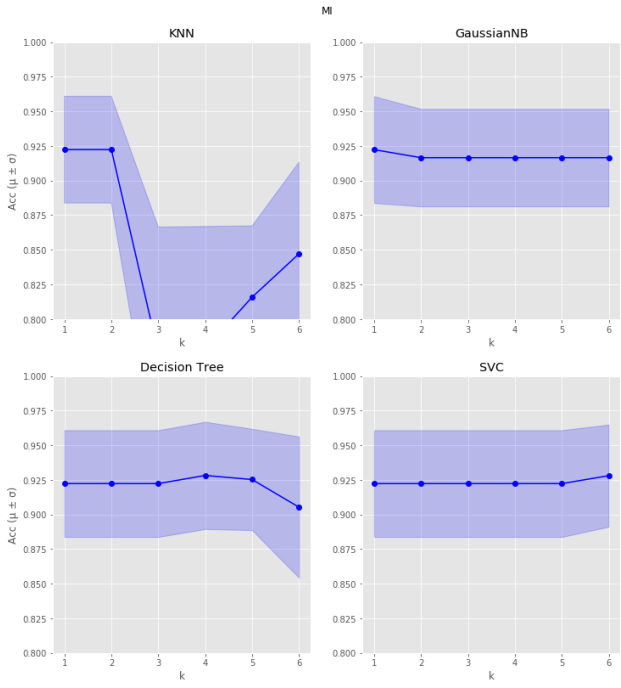
A quick remark on the notations and parameters of the problem: $d = 30$ refers to the dimensionality of the search problem, i.e. dimension of the feature space; $m = 10$ is the total number of subsets acquired from the ten cross-validation folds. The feature subset cardinality is restricted to 6 features ($k = 6$) for all m subsets.

Regarding the values reported on Tables 4.1 and 4.2: S_{iMax} is the subset of features with maximum intersection from all the feature subsets of the cross-validation and r refers to its cardinality, i.e. the size of the subset intersect. k_{best} is the feature number that *first* reached the highest performance score on each fold of the cross-validation. Namely, the lowest value of k with the highest performance score on each fold was examined, as in some folds the highest score was repeated for several k values. s is the number of folds with each possible k_{best} number of features. Finally, a best final subset of features S_{best} is selected according to the criteria presented in Section 3.2.3.

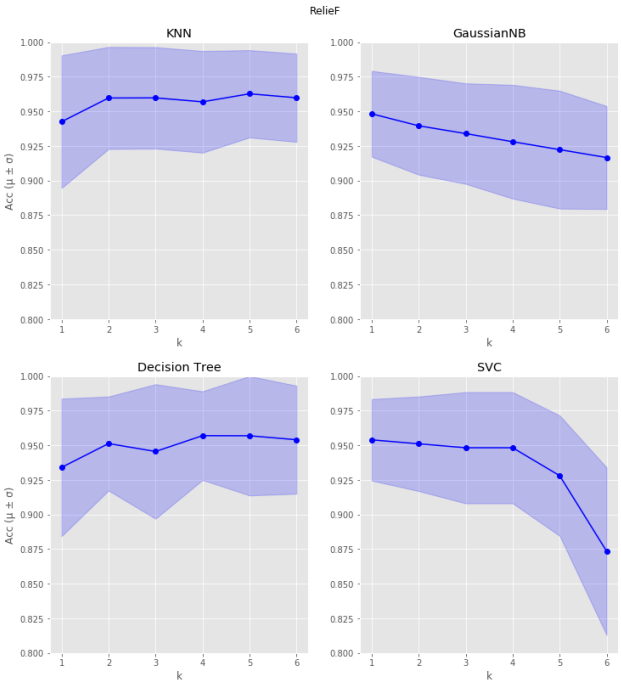


(a) ANOVA, Analysis of Variance

(b) Chi-squared



(c) MI, Mutual Information



(d) Relief Algorithm

Figure 4.1: Feature selection: univariate filter accuracy scores.
Parametric methods: ANOVA (Analysis of Variance) (a). Non-parametric methods: Chi-Squared (b), MI (Mutual Information) (c) and Relief algorithm (d). k, number of features (X-axis); acc ($\mu \pm \sigma$), accuracy cross-validation scores reported with μ – mean – and σ – standard deviation – values (Y-axis); KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

4.1.1 Univariate Filter Feature Selection

The honestly cross-validated accuracy results of the four univariate filter methods introduced in Section 3.2.1 are shown in Figure 4.1.

The experimental results follow a quite similar tendency for the case of chi-squared (b) and MI (c) filters. The average accuracy classification scores for all scenarios are generally high, that is, over 0.9 except some cases when using the KNN and SVM classifiers. The KNN classifier's accuracy dropped when selecting features subsets of more than two features with the chi-squared and ANOVA univariate filters. On the other hand, a similar but less dramatic downward trend occurs to the SVM using the Relief algorithm to select feature subsets of more than 5 features. Nevertheless, its tendency in the long-term is unknown, as feature subsets over 6 features are not shown, whereas an accuracy over 0.9 was never reached again by the KNN classifier when selecting more than two features.

ANOVA and Relief filters yield the best performance results for all the classifiers, but the accuracy in the former showed a decrease with higher number of features if Naïve Bayes or SVM classifiers were used. In all, the results for all methods using ANOVA and Relief filters rounded 0.95, whereas chi^2 and MI did generally not reach 0.925 accuracy values.

Another relevant fact of univariate filter selection is that the top 4 features from the ranks resulting in the four methods were consistent, i.e. with high stability in the resulting feature subsets, when applied in the whole dataset, as shown in Figure A.5.

More information with regard to the features selected by the filter methods are reported in Table 4.1.

- S_{iMax}. The ANOVA, chi^2 and Relief methods show more consistency on the ranked features across the cross-validation folds, with almost or complete accordance on the rank order. On the contrary, there is low (i.e. $r = 2$) or no consistency when applying the MI filter.

- k_{best}.The variability of the feature number that first reached the highest performance score on each fold, i.e. k_{best}, depends on the classifier and the filter metric used. In all univariate filter methods, great variability exists when using the Decision Tree classifier. Otherwise, total agreement in all folds of the k_{best} number rarely occurs (e.g. k = 1, s = 10), only in the NB classifier using chi^2 and in SVM using ANOVA and Relief metrics. In such cases, it only occurred with only one feature subsets, namely the GCIPL layer in the 3-to 6-mm ring.

Filter, d=30, m=10, k=6	r	S_{iMax} **(subset intersect)**	k_{best} **(first best k), s (sum across fold)**	**final sequence criterion**	S_{best} **(final subset)**
ANOVA					
KNN	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl_1 mm	k = 1, s = 4 k = 2, s = 4 k = 3, s = 2	S^*_{minE}	GCIPL 3 6 mm
NB	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl_1 mm	k = 1, s = 9 k = 4, s = 1	S^*_{rank}	GCIPL 3 6 mm
DecTree	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl_1 mm	k = 1, s = 3 k = 2, s = 3 k = 3, s = 2 k = 4, s = 1 k = 5, s = 1	S^*_{minE}	GCIPL 3 6 mm GCIPL 3 6 mm, GCIPL 6 mm
SVM	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl_1 mm	k = 1, s = 10	S^*_{rank}	GCIPL 3 6 mm
chi^2					
KNN	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl 1 mm	k = 1, s = 9 k = 6, s = 1	S^*_{rank}	GCIPL 3 6 mm
NB	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl 1 mm	k = 1, s = 10	S^*_{rank}	GCIPL 3 6 mm
DecTree	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl 1 mm	k = 1, s = 4 k = 3, s = 1 k = 4, s = 1 k = 5, s = 3 k = 6, s = 1	S^*_{minE}	GCIPL 3 6 mm
SVM	5	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, rnfl 1 mm	k = 1, s = 5 k = 5, s = 5	S^*_{minE}	GCIPL 3 6 mm
MI					
KNN	2	GCIPL 3 6 mm, GCIPL 6 mm	k = 1, s = 9 k = 6, s = 1	S^*_{rank}	GCIPL 3 6 mm
NB	0		k = 1, s = 9 k = 2, s = 1	S^*_{minE}	GCIPL 6 mm GCIPL 3 6 mm
DecTree	2	GCIPL 3 6 mm, GCIPL 6 mm	k = 1, s = 7 k = 3, s = 1 k = 4, s = 1 k = 5, s = 1	S^*_{minE}	GCIPL 3 6 mm
SVM	0		k = 1, s = 5 k = 2, s = 1 k = 5, s = 2 k = 6, s = 2	S^*_{minE}	GCIPL 6 mm, GCIPL 3 6 mm

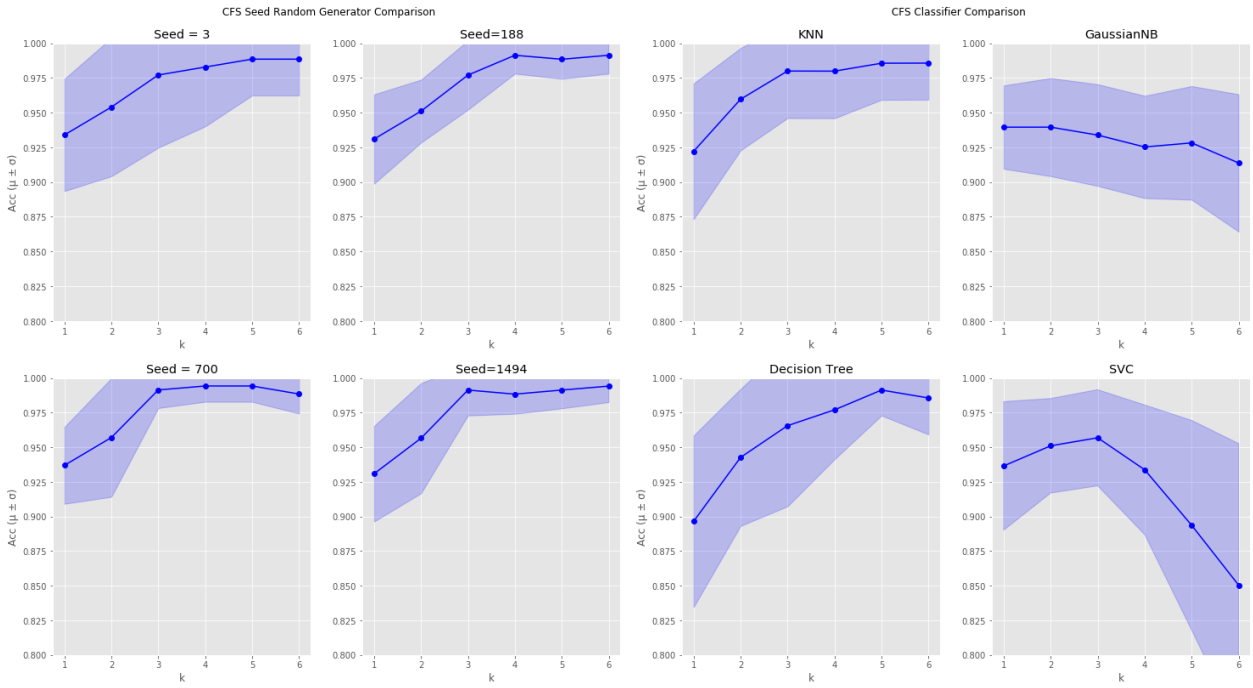
Relief					
KNN	6	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm,GCIPL 3 mm, retina 1 3 mm, retina 3 mm	k = 1, s = 2 k = 2, s = 4 k = 3, s = 2 k = 5, s = 2	S^*_{minE}	GCIPL 3 6 mm GCIPL 3 6 mm, GCIPL 6 mm GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, retina 1 3 mm
NB	6	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, retina 1 3 mm, retina 3 mm	k = 1, s = 9 k = 4, s = 1	S^*_{rank}	GCIPL 3 6 mm
DecTree	6	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, retina 1 3 mm, retina 3 mm	k = 1, s = 2 k = 2, s = 3 k = 3, s = 1 k = 4, s = 2 k = 5, s = 1 k = 6, s = 1	S^*_{minE}	GCIPL 3 6 GCIPL 3 6 mm, GCIPL 6 mm GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm
SVM	6	GCIPL 3 6 mm, GCIPL 6 mm, GCIPL 1 3 mm, GCIPL 3 mm, retina 1 3 mm, retina 3 mm	k = 1, s = 10	S^*_{rank}	GCIPL 3 6 mm

Table 4.1: Univariate filters FS.
ANOVA, Analysis of Variance; chi^2, Chi-square; MI, Mutual Information; d, feature space dimension ; m, total number of subsets in cross-validation; k, maximum feature subset size; r, maximum size subset intersect across cros-validation folds; S_{iMax}, features in the maximum intersect subset; k_{best}, number of features with best performance score; s, sum of folds with each k_{best}; S_{best}, selected final subset; S^*_{rank}, consistency selection criterion; S^*_{minE}, criterion for unconsistency results; KNN, K-Nearest Neighbors Classifier; NB, Naïve Bayes Classifier; DecTree, Decision Tree; SVM, Support-Vector Machine; GCIPL, ganglion cell-inner plexiform layer;rnfl, retinal nerve fiber layer; retina, total thickness value of retinal layers.

The degree of variability of k_{best} affects on the selection of a S_{best} subset. Nevertheless, the selected features show few differences across univariate filter metrics as well as across cross-validation folds, both criteria resulted in similar results.

- As a result, one finding common in all comparisons – despite some alterations in the ranking order – is that the outer regions of the GCIPL region are the most relevant to discriminate between the patient and control groups (3-to 6-mm ring and the 6 mm disc).

4.1.2 Multivariate Filter Feature Selection



(a) Random seed using KNN classifier **(b)** Classifier algorithms

Figure 4.2: Feature selection: multivariate filter accuracy scores.
Experimental runs: random seed generator (a) and classifier algorithm (b) comparisons. k, number of features (X-axis); Acc ($\mu \pm \sigma$), accuracy cross-validation scores reported with μ – mean – and σ – standard deviation – values (Y-axis); KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

Similar to the experimental runs using univariate filter methods, four different classifiers were used to evaluate performance in classification with multivariate feature selection. Additionally, to account for another potential source of variability, the experiment was run using different random seeds to create the folds in the cross-validation. This parameter was set constant in the rest of comparisons (seed = 1313).

CFS d**=30,** m**=10,** k**=6**	r	S_{iMax} **(subset intersect)**	k_{best} **(optimal feature number),** s **(sum across fold)**	**final sequence criterion**	**S_best (final subset)**
KNN	0		k = 2, s = 3 k = 3, s = 5 k = 4, s = 1 k = 5, s = 1	S*_minE	GCIPL 3 6 mm, GCIPL 6 mm GCIPL 3 6 mm, GCIPL 6 mm, retina 6 mm GCIPL 3 6 mm, GCIPL 6 mm, retina 1 3 mm, retina 3 6 mm
NB	0		k = 1, s = 9 k = 2, s=1	S*_minE	GCIPL 3 6 mm GCIPL 6 mm
DecTree	0		k = 2, s = 2 k = 3, s = 4 k = 4, s = 1 k = 5, s = 3	S*_minE	GCIPL 6 mm, GCIPL 3 6 mm GCIPL 6 mm, GCIPL 3 6 mm, retina 6 mm GCIPL 6 mm, GCIPL 3 6 mm, retina 3 6 mm GCIPL 6 mm, GCIPL 3 6 mm, retina 6 mm, retina 3 6 mm GCIPL 6 mm, GCIPL 3 6 mm, GCIPL 3 mm, retina 1 3 mm, retina 6mm
SVM	0		k = 1, s = 5 k = 2, s = 3 k = 3, s = 2	S*_minE	GCIPL 6 mm GCIPL 6 mm, GCIPL 3 6 mm, retina 6 mm

Table 4.2: Multivariate filter FS: CFS.
CFS, Correlation-based Feature Selection; d, feature space dimension ; m, total number of subsets in cross-validation; k, maximum feature subset size; r, maximum size subset intersect across cros-validation folds; S_{iMax}, features in the maximum intersect subset; k_{best}, number of features with best performance score; s, sum of folds with each k_{best}; S_{best}, selected final subset; S^*_{rank}, consistency selection criterion; S^*_{minE}, criterion for unconsistency results; KNN, K-Nearest Neighbors Classifier; NB, Naïve Bayes Classifier; DecTree, Decision Tree; SVM, Support-Vector Machine; GCIPL, ganglion cell-inner plexiform layer; retina, total thickness value of retinal layers.

The following conclusion can be drawn:

- Fig. 4.2 (a), random seed generator comparison. Although the accuracy scores are not the same, trends remain similar when folds are generated with different random seeds, in comparison with the differences that appear between alternative classification algorithms, as shown in (b).

- Fig. 4.2 (b), classifier algorithm comparison. The highest classification scores are reached when using KNN and Naïve Bayes classifiers. The CFS multivariate filter shows the greatest variability in the ranking order of subsets among all the filter

methods presented in this work, as the rank order differs across folds and, thus, no subset interesect could be determined from this filter method (see S_{iMax} in Table 4.2). Therefore, all the subsets that reached highest performance scores were reported as S_{best}, according to the S_{minE} criteria presented in Section 3.2.3.

- Table 4.2 suggests the feature number has less impact on the classification performance, as the highest scores were obtained with different k values and the selection of a k_{best} was not possible.

  Nevertheless, the final feature subsets are consistent with previous results: outer regions of the GCIPL layer are selected. Additionally, whole retina thickness and GCIPL in the outer and inner regions – in KNN and Decision Tree classifiers – also yield the highest scores, whereas NB and SVC showed lower feature number.

4.1.3 Wrapper Feature Selection

Greedy Search: Sequential forward/backward selection

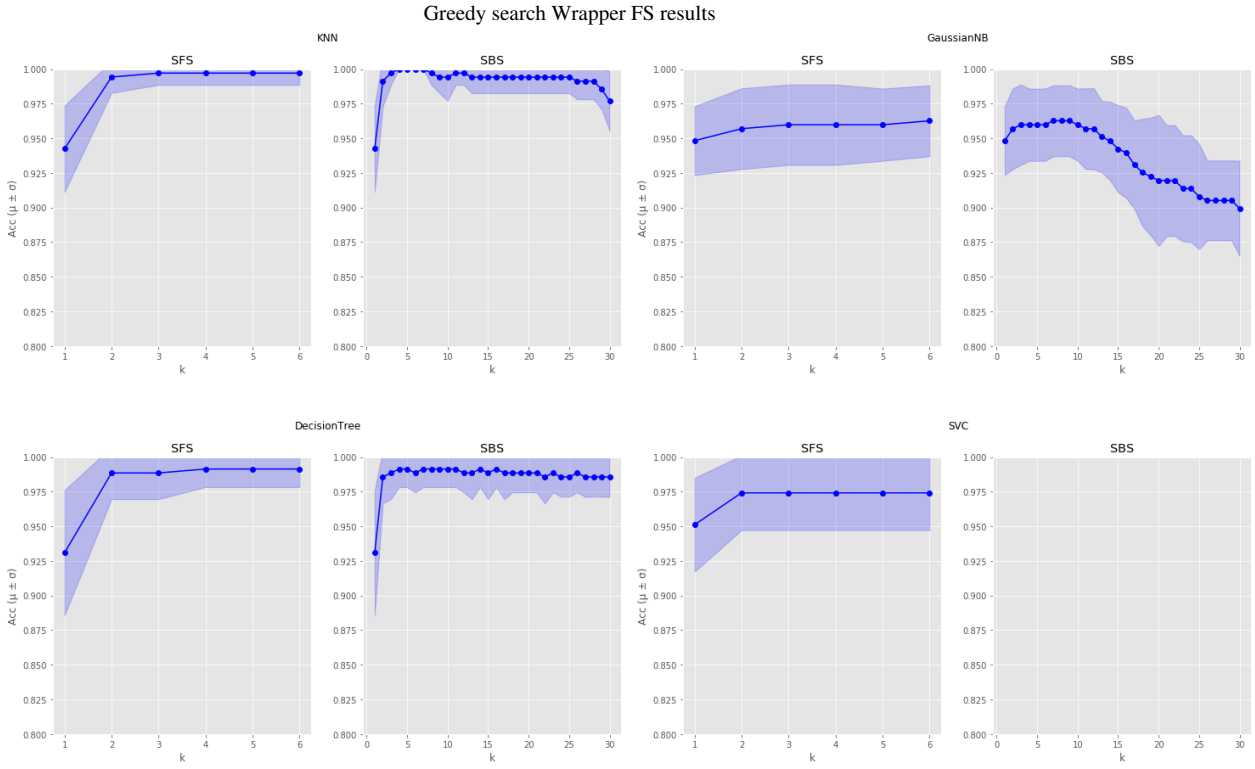
The results obtained using the wrapper feature selection methods presented in Section 3.2.2 are shown in Figure 4.3.

Three performance score metrics are used, in order to gain deeper insight of the classification within the patient class: (a) accuracy, (b) recall and (c) precision. The last two are included due to the importance of the patient group, which is a minority group in the data set, so the global results in terms of accuracy can be misleading.

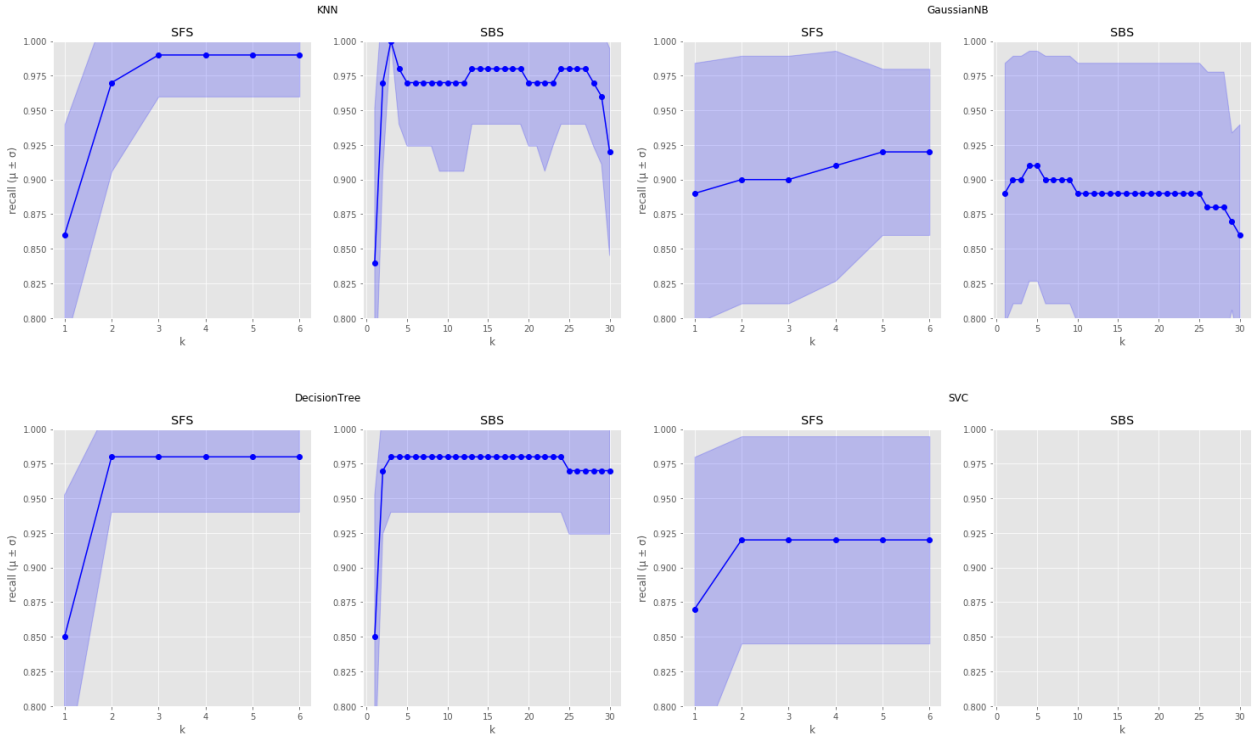
KNN and Decision Tree classifiers yield significantly higher scores than the other two in all three metrics and in both sequential forward (SFS) and backward (SBS) selection. Furthermore, the results using SVM in sequential backward selection do not reach 0.8 (see Fig. 4.3 (c)). Nevertheless, the values of all the metrics are quite high, i.e. within [0.8, 1] range, and generally over 0.9 in all comparisons.

As expected, differences between accuracy and recall-precision values were found. Recall had lower mean and higher standard deviation values than the global accuracy scores in all four classifiers. A similar decrease of performance occurs in precision when using Decision Tree and Naïve Bayes classifiers.

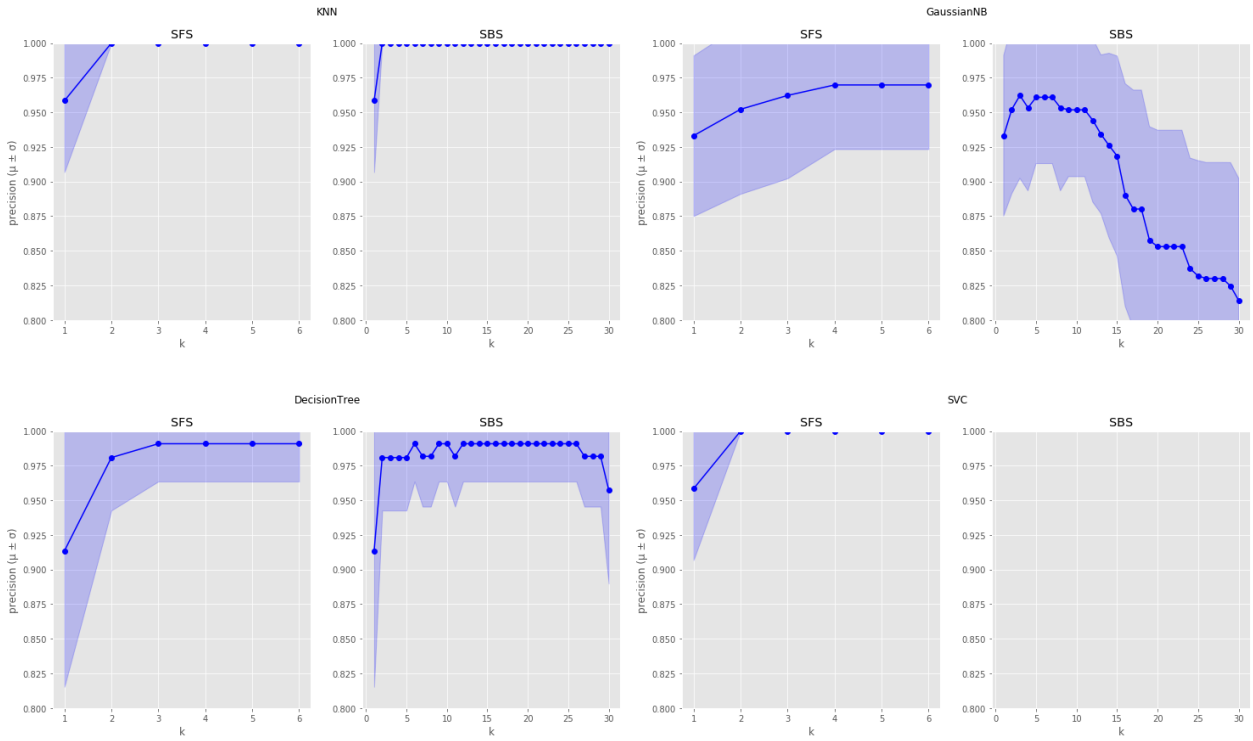
Intuitively, precision is the ability of the classifier to correctly assign the positive labels, and recall is the ability of the classifier to find all the positive samples. From the figures, both abilities appear to constrained.



(a) Accuracy



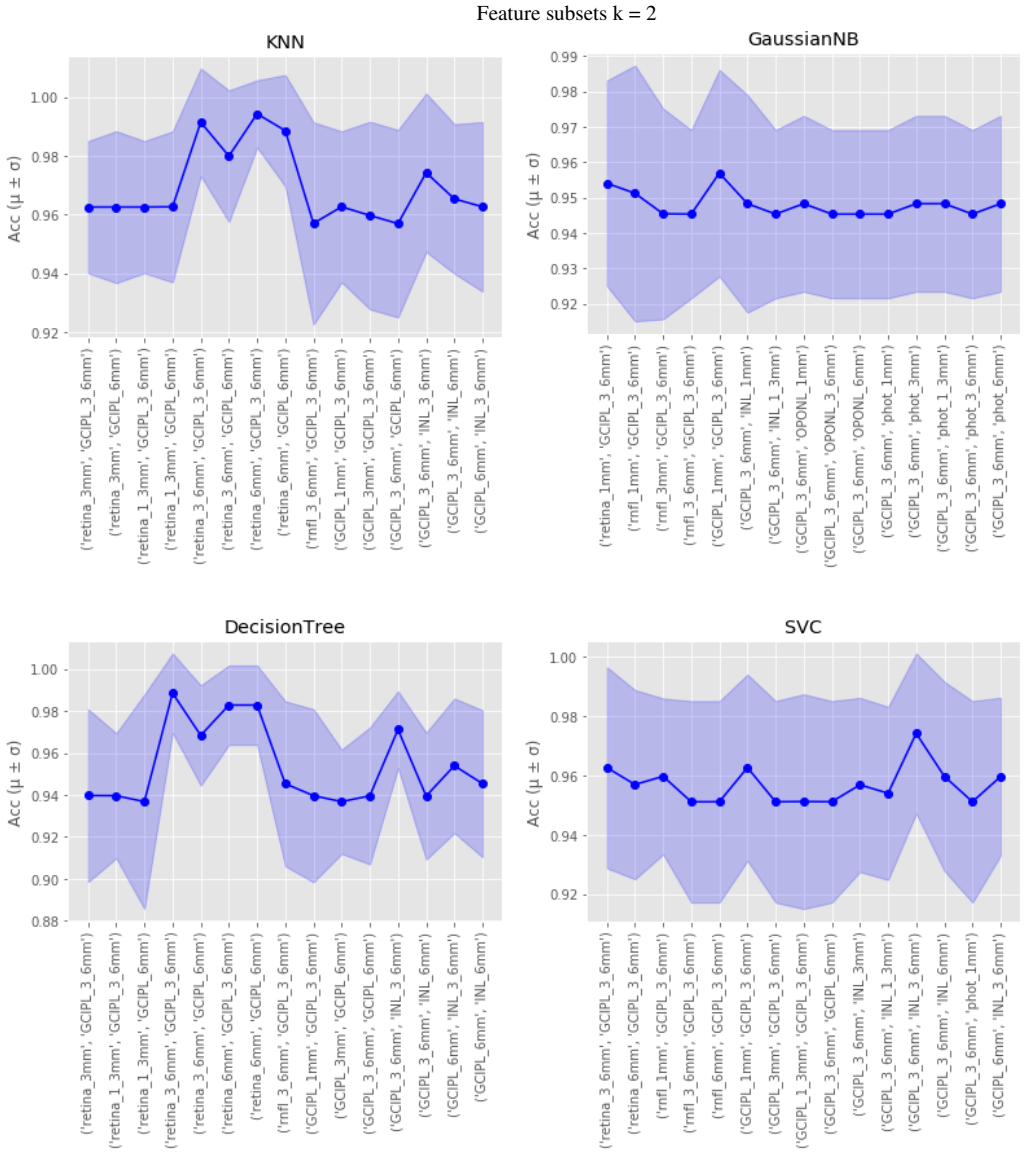
(b) Recall scores



(c) Precision

Figure 4.3: Feature selection: wrapper greedy search.
Performance plots Sequential forward/backward selection performance plots of: (a) accuracy, (b) recall and (d) precision. SFS, sequential k, number of features (X-axis); Acc, accuracy; μ, mean value; σ, standard deviation; KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

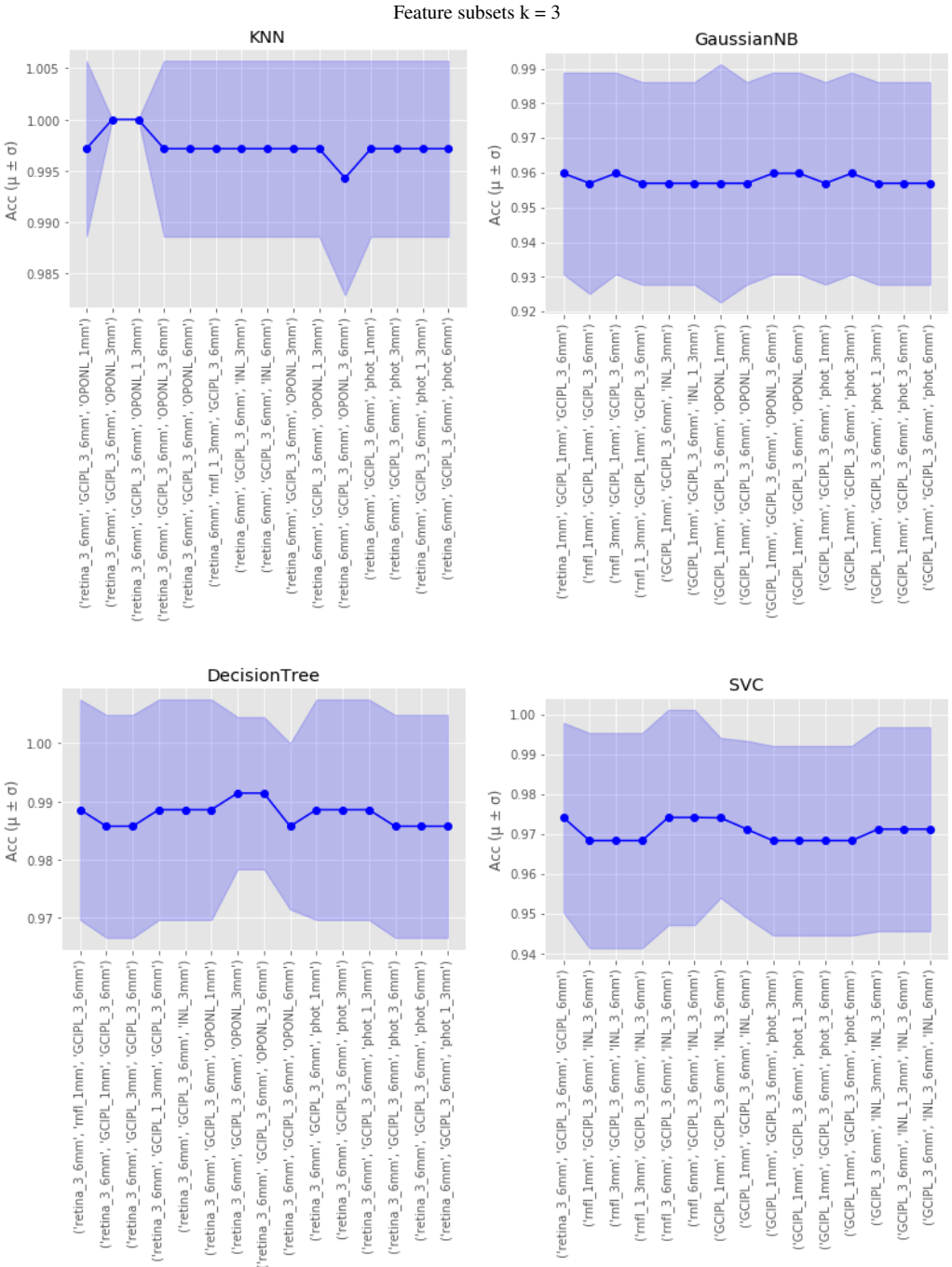
Bounded Exhaustive feature selection: k = 2,3



(a) k = 2

Exhaustive feature selection evaluates the performance of all possible combinations of features. Due to the dimensionality of the problem, mean and standard deviation accuracy scores were calculated for all the possible combinations of feature subsets constrained to two or three features.

As the number of possible combinations is certainly large, only the 15 subsets with highest scores were selected and presented in Figure 4.4 for subsets of k = 2 (a) and k = 3 (b) features. In the figures, it can be observed that the mean values across feature subsets of 2 features show greater variability than those of 3 features in all classifiers.



(b) k = 3

Figure 4.4: Feature selection: bounded exhaustive search.
Wrapper FS using bounded exhaustive feature seleciton with k = 2 (a) and k = 3 (b), left and right figures in each subfigure, respectively. k, number of features (X-axis); recall ($\mu \pm \sigma$), recall cross-validation scores reported with μ – mean – and σ – standard deviation – values (Y-axis); KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

KNN and Decision Tree classifiers provide the results with highest mean values and lower standard variation. With two features (Fig. 4.4 (a)), three subsets yield accuracy over 0.98 only with KNN and Decision Tree classifiers. The corresponding subsets match in both classifiers and are: 3-to 6-mm ring in GCIPL and retina, 6 mm disc in GCIPL and retina, and GCIPL 3 6 mm and retina 6 mm subsets.

Once again, the method is consistent with the results of the filter feature selection methods, i.e. outer regions of GCIPL layer and retina are the most informative and relevant in this classification problem.

4.2 Dealing with imbalanced data

The purpose of this section is to evaluate the classification ability of retinal thinning accounting for the bias towards the minority class that occurs due to the difference of sample size between the two classes.

In this context, two scenarios are presented in Tables 4.3 and 4.4: the 10-fold cross-validation scores of the original data set (referred as ORIGINAL in the tables) and the data set with equal proportion between both class in the train set generated with the SMOTE algorithm. Additionally, the results are reported with several metrics. Recall and precision scores, due to the relevance of the patient class, as well as the arithmetic and geometric mean of recalls to obtain a trade-off of the classification ability of both classes. The former are presented in Section 4.2.1 and 4.2.2, respectively.

4.2.1 Imbalanced score metrics

The results of the 10-fold cross-validation using exposed four classifiers are presented on Table 4.3. Regarding the statistics computed from the confusion matrix given the two possible scenarios, it can be concluded that:

- All metrics yield higher mean score values in the data set with equal proportion of both classes in training than the original dataset in the KNN and SVM classifiers.
- Furthermore, recall and precision scores are the ones of best improvement. From the table, it can be noticed that the minority class is specially affected in the SVM classifier. Due to the lack of incorrectly or correctly labelled samples in the test

sample, it resulted in a zero value across all folds in the cross validation, which will be referred in the following as the "zero division" problem. It can be seen that after applying the SMOTE technique, the results of this classifier are similar to the ones of the other three classifier algorithms. The same situation applies to the precision score when using KNN classifier.

- Finally, from the highest values which are highlighted in the tables, the KNN is the one with best performance and, in the SMOTE data set, in unison across all shown scores.

ORIGINAL	**accuracy**	**recall**	**precision**
KNN	**0.9767 ± 0.0176**	0.9181 ± 0.0657	1 ± 0
NB	0.904 ± 0.049	0.8804 ± 0.0931	0.8108 ± 0.1037
DecTree	0.9736 ± 0.0334	**0.9479 ± 0.1011**	**0.9578 ± 0.0685**
SVM	0.7094 ± 0.0479	0 ± 0	0 ± 0

SMOTE	**accuracy**	**recall**	**precision**
KNN	**0.9883 ± 0.0143**	**0.9764 ± 0.0473**	**0.9823 ± 0.0358**
NB	0.8924 ± 0.0471	0.8672 ± 0.0989	0.7876 ± 0.0956
DecTree	0.9708 ± 0.0473	0.9444 ± 0.1139	0.9453 ± 0.0854
SVM	0.9186 ± 0.0451	0.9006 ± 0.1045	0.8453 ± 0.1092

Table 4.3: Classification non-balanced scores: accuracy; recall and precision of the minority class (patient group).
10-fold cross-validation scores obtained in the dataset with original class proportions (ORIGINAL) and in an equal proportion of training set (SMOTE) data sets. KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

4.2.2 Balanced score metrics

The score metrics used in Table 4.4 are referred as balanced classification metrics, because they incorporate the classification ability of both classes. As a result, the metrics are less biased to disproportions between classes.

- Similar to the results in the previous table, the scores of all metrics are improved in the data set with equal training proportion among classes.
- Moreover, imbalance affects differently to each classification algorithm. The "zero division" problem has greater impact on the SVM classifier before SMOTE because then, the data is not representative enough of the minority group during training.

- Decision Tree yields the highest performance scores for the balanced classification scores. In the data set with equal training proportion, it is the KNN once again the classifier with best performance.
- A final remark, which is common in the non-balanced metrics of Table 4.3, is that even when there is no "zero proportion" problem in the SVM classifier, its scores are significantly lower than those of the other three classifiers. This leads to the idea of the SVM as not optimal for this problem in particular.

ORIGINAL	**recall_maj**	**recall_min**	**recall_A_mean**	**recall_G_mean**
KNN	1 ± 0	0.9181 ± 0.0657	0.9591 ± 0.0329	0.9576 ± 0.0346
NB	0.9135 ± 0.0527	0.8804 ± 0.0931	0.8969 ± 0.0598	0.8955 ± 0.061
DecTree	$\mathbf{0.9843 \pm 0.0263}$	$\mathbf{0.9479 \pm 0.1011}$	$\mathbf{0.9661 \pm 0.0526}$	$\mathbf{0.9643 \pm 0.069}$
SVM	1 ± 0	0 ± 0	0.5 ± 0	0 ± 0

SMOTE	**recall_maj**	**recall_min**	**recall_A_mean**	**recall_G_mean**
KNN	$\mathbf{0.9916 \pm 0.0169}$	$\mathbf{0.9764 \pm 0.0473}$	$\mathbf{0.9840 \pm 0.0231}$	$\mathbf{0.9836 \pm 0.0238}$
NB	0.9010 ± 0.0508	0.8672 ± 0.0989	0.8841 ± 0.0576	0.8822 ± 0.0587
DecTree	0.9803 ± 0.0317	0.9444 ± 0.1139	0.9624 ± 0.0671	0.9607 ± 0.0706
SVM	0.9248 ± 0.0584	0.9006 ± 0.1045	0.9127 ± 0.0571	0.9103 ± 0.0589

Table 4.4: Classification balanced scores: majority/minority recall, arithmetic/geometric recall mean.
10-fold cross-validation scores obtained in the dataset with original class proportions (ORIGINAL) and in an equal proportion of training set (SMOTE) data sets. KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

4.3 Dealing with effect of age

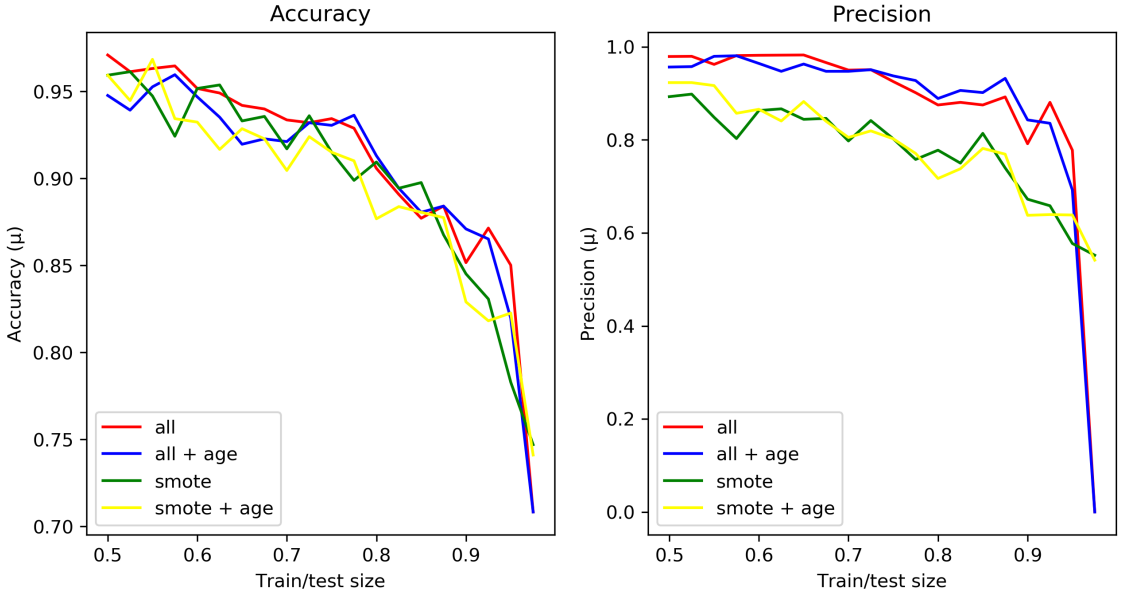
Previous research of retinal thinning revealed the impact of age in macular degeneration. Lower thickness values are expected only due to the mere fact of growing older. Therefore, this additional variable is important when evaluating the prediction ability of the retina.

Several classification scores were computed including and excluding the age variable. Furthermore, the original data set and the one generated with the SMOTE technique were used for comparison. The results are shown in Fig. 4.5.

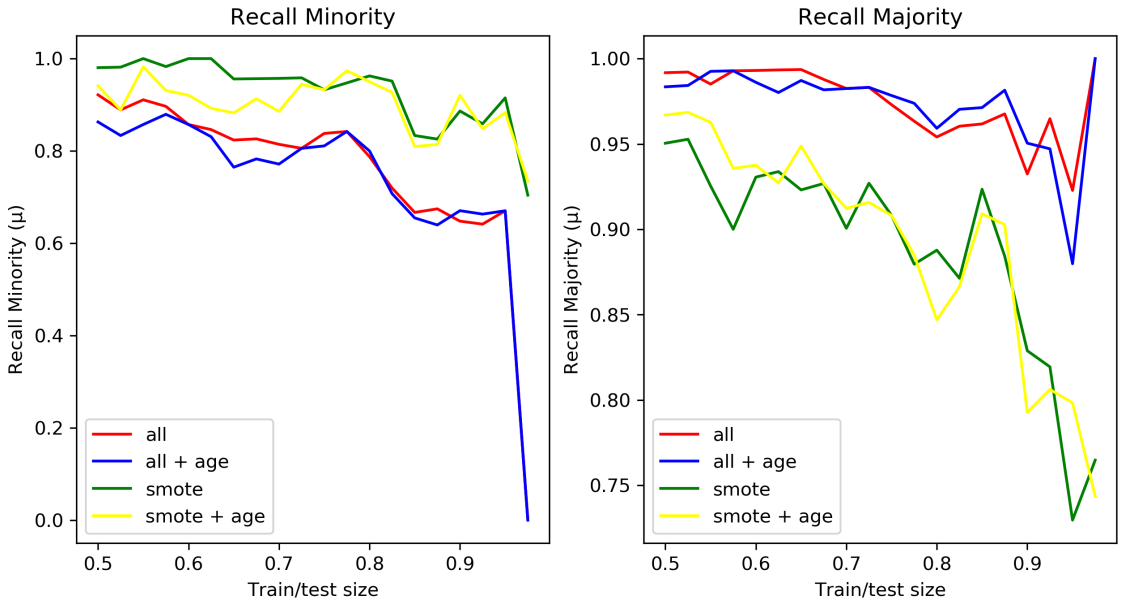
The scores were acquired from partitions of different size within a train/test split scheme, that is, using a proportion of the data for training the model and testing on the remaining part. The plots were computed starting with 0.5 of the sample used for training and the other 0.5 for testing, with increments of 0.025, i.e. in total 20 experimental runs for each score are shown on each figure. The classification algorithm selected was the KNN classifier.

On the one hand, the imbalance effect is evaluated. For this purpose, the original data set (red/blue) is compared with the SMOTE data set (green/yellow). Accuracy and recall of the minority group improved when SMOTE is applied, whereas a slight drop in precision and notable in recall of the majority group are observed. Consequently, the imbalance effect is notably better accounted for with the arithmetic and geometric mean of the results from the two classes, which are shown in (c), as the gap between the original and the SMOTE data sets is wider than the global results shown in (a) (the blue/red and green/yellow lines respectively).

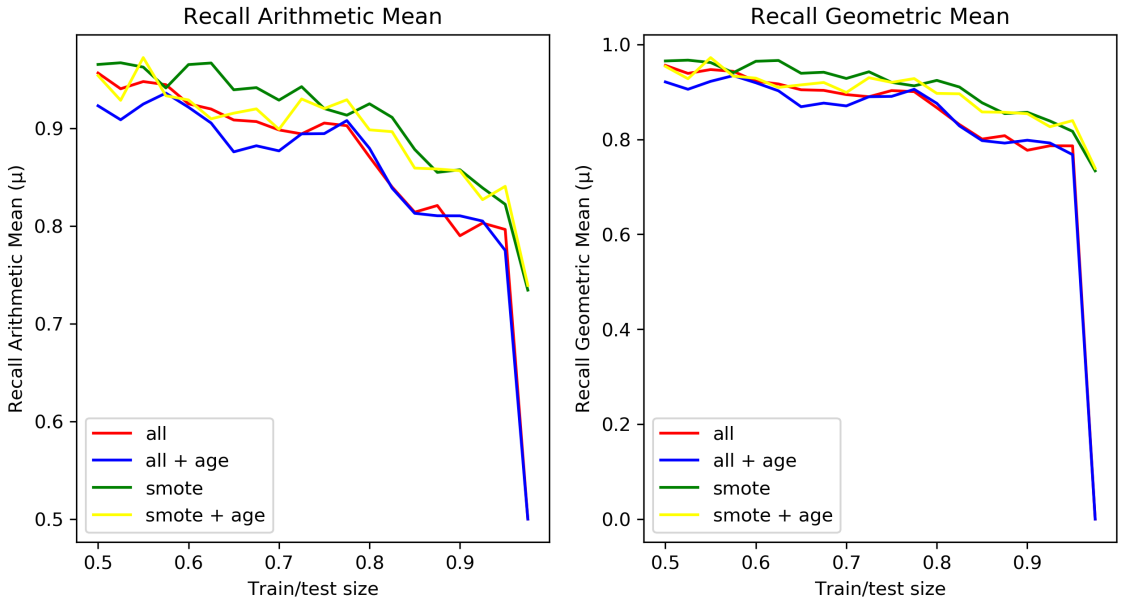
On the other hand, the classification including age is examined. In this scenario, the results within the two possible data sets including and excluding the age are compared, i.e. red/green with respect to blue/yellow. First, it is true that the differences when including the age in the classification ability are not strong and fluctuate throughout the train/test proportions. Overall, the classification ability decreases when age is included in both data sets.



(a) Accuracy and Precision



(b) Minority/majority class recall



(c) Arithmetic/geometric mean recall

Figure 4.5: Train/test split classification results.
Dealing with effect of imbalance and inclusion of age. Use of the K-Nearest Neighbors classifier, different proportion size (Train/test size) and scores: accuracy and precision (a), recall of the minority/majority classes (b) and arithmetic/geometric mean recall values (c). μ, mean value.

5. CHAPTER

Discussion

5.1 Preliminary analysis of the data

This work aim to evaluate the classification ability of 30 variables of the retina extracted using a commonly used imaging technique in ophthalmology, i.e. OCT, and the relevance of which variables better discriminate patients suffering Parkinson's Disease from healthy control data.

For this purpose, supervised learning techniques have been applied to the data set. Yet, the first step consisted of an exploratory analysis of the data set, which already revealed some patterns of interest.

Firstly, box-plots were used to search for potential outliers according to the $3\sigma - rule$, which were imputed with the median value. Outliers were found mainly in control data, within the 1-mm disc region. These plots also revealed the data distribution conditioned to each class group, shown in Fig. A.1. From them, one can easily notice greater differences among classes on the *GCIPL* layer with respect to the rest retinal layers, being the greatest a $20 - \mu m$ gap between median values in the outer regions of this layer.

Secondly, all variables showed to follow a normal distribution when looking to the histograms of each class group. Only for five variables the null hypothesis of D'Angostino and Pearson's normality test were discarded: 1-mm and 3-mm discs and 1-to 3-mm ring in INL layer and 3-to 6-mm ring in rnfl in controls; and the last, 1-mm disc in GCIPL, in both control and patient data. Consequently, these variables could not be included in

ANOVA filter feature selection, which is a parametric statistical test.

Finally, variables were grouped together according to higher Pearson correlation values (correlation plot of Fig. 3.1) and therefor, the ones from a same retinal layer were grouped. Moreover, correlation values can be examined in Fig. A.4 for control and patient data. In the latter, some variables within the same layer showed intra-layer correlation values over 0.95. The total thickness value of the retina showed the greatest inter-layer correlation and mainly with *GCIPL*, rounding 0.8 values. The *INL* and the *OPONL* layers followed, with correlation values around 0.65. Next, values around 0.4 were found with the $rnfl$ layer and finally, low correlation with the photorreceptor layer, i.e. around 0.1. Similarly, the same order of correlation with total retinal thickness applied in controls, although two main differences were found: in *GCIPL* and photorreceptor layers. The correlation of the former were lower than the ones of patients, i.e. roughly 0.75. On the other hand, the photorreceptor layer was more correlated with the retina, that is, in the same degree as the $rfnl$ layer, with values close to 0.4.

5.2 Classification analyses

On the other hand, the goal of the analysis was to determine the features which carry more information and relevancy on testing the hypothesis of retinal thinning as a potential biomarker of Parkinson's disease, and to evaluate to which extend this hypothesis is true. For this purpose, retinal measurements of a hundred patients suffering Parkinson's disease and 248 healthy controls were compared using several supervised learning techniques. In this section, factual support to the questions addressed at the beginning of this work is provided.

To which extend can retinal thickness values be a valid biomarker of the disease?

There is proof in this work that retinal thinning allows to discriminate patients with Parkinson's disease from healthy controls. Supervised classification with four different classification algorithms revealed high honestly cross-validated accuracy scores, i.e. over 0.8, in all conducted experiments. The classification ability of these measurements was also examined in terms of precision and recall, due to the clinical relevancy of the patient group. However, this last two metrics encountered a problem caused by the lower proportion of positive samples in the data set, which generated 0 score values. In scikit learn it is referred as the "zero division" problem, and this was the motivation to use oversampling techniques, in particular SMOTE, to deal with imbalanced data. After SMOTE was

applied, the were again high performance scores, specially in KNN and Decision Tree classifiers, with values over 0.9 (see Tables 4.3 and 4.4). Finally, the trade-off of performance of both classes was computed through the mean of recall values estimated from each class group, to obtain the arithmetic and geometric mean. These metrics, which are *balanced* score metrics, were again high in all the classification algorithms.

Which are the most relevant layers and regions in the early diagnosis of the disease?

Measurements extracted from the GCIPL layer were evidently the ones with more relevance and information in the prediction of the disease, as concluded from the results of several feature selection techniques. The results acquired on the face of feature selection were higher than those including the whole data set (see Figures 4.1 and 4.2). Moreover, some interesting facts were discovered by looking into the subsets that reached the highest scores on each cross-validation fold. The outer regions of GCIPL layer, namely the 3-to 6-mm ring and the 6-mm disc, were selected by all univariate and multivariate filter methods. Multivariate filters also selected the whole retina thickness in the same regions. This nuance of the multivariate subsets with the univariate filter ranks can be expected and explained, due to the high correlation with the GCIPL layer with the whole retina thickness.

On the other hand, there was no evidence of a given feature subset cardinality to yield better performance, which can be given the high inter-feature correlation of some of the most relevant features. Consequently, the highest performance score was frequently reached within each validation fold for several feature numbers. For this reason, the subsets with minimal feature number are the ones presented on Tables 4.1 and 4.2, which are based on searching the minimal feature number which yield the highest performance score on each fold, and afterwards comparing the consistency across cross-validation folds to select the best subset. Two options applied to select the best subset: the intersection across folds if there was enough consistency, or returning all the subsets that reached the highest score in the cross-validation. Namely, sometimes there was no intersect between subsets of the cross-validation folds and the highest performance value was reached several times with different cross-validation fold subsets.

Low variability of performance given the number of features selected was observed in sequential forward/backward selection results of the wrapper, as shown in Fig. 4.3. Accuracy, precision and recall scores showed no significant differences when k, feature subset number, was respectively increased/reduced. The only exception was the Naïve Bayes classifier, in which the performance dropped with higher k number and thus, a greater

probability of feature redundancy. The performance was still very high for all shown scores. Besides, subsets consisting on GCIPL and retina in the outer regions reached the highest accuracy among all possible combinations of subsets with two features in presented four classification algorithms when bounded exhaustive feature selection was applied. In three feature combinations, no subsets within the 15 selected best subsets outperformed the others notably, only in KNN and Decision Tree classifiers. In such cases, the subsets consisted again of GCIPL and retina in the outer ring, with addition of the OPONL layer. These conclusions are drawn from the bounded exhaustive search shown in Fig. 4.4.

All in all, supervised learning methods respond to the two questions raised at the beginning of this work to some extend. The retinal thinning, specially of some layer and regions, has the ability to perform a rather accurate discrimination of patients with Parkinson's disease. From the presented classification algorithms, the KNN classifier showed the highest performance, while the SVM obtained the lowest, in this particular classification problem. The presented results aid for a better understanding of the relevance of the retina for the diagnosis of Parkinson's disease. It was indeed possible through the assessment of the classification ability of a data set consisting of measurements from all retinal layers and regions presented in Section 2.5 (shown in Fig. 2.4), while accounting with the high inter-feature correlations, as well of the performance of some selected features to produce more accurate results.

6. CHAPTER

Conclusions and future work

This work addresses a complex problem, on the face of a disease with different phenotypes and no defined protocol for its early diagnosis. [Bodis Wollner and Yahr, 1978] claimed evidence of visual dysfunction in Parkinson's disease already in 1978, and ever since related research involved in the anatomy of the retina has been conducted. To this date and year, OCT is the imaging technique most widely used to acquire measurements of the retina. This work used OCT to compute up to 30 variables which represent the mean thickness values in μm of five retinal layers, in different ring and disc regions of the back of the eye, i.e. funduscopic images.

Several state-of-the-art techniques are used for this binary supervised learning problem. Their findings help to consolidate the hypothesis of retinal thinning caused by the disease, which as a result can be a valid biomarker of the disease, in particular of the GCIPL layer, i.e. ganglion-cell inner-plexiform layer. The results not only discriminate the patient group with high accuracy, precision and recall scores, but also show an improvement of classification performance when feature subsets with lower cardinality are used, including only GCIPL and total retina thickness values in the outer region of the eye.

This work offers a study of hidden patterns of data recorded from the retina of a considerable number of subjects and using a variety of methods. The latter mainly include feature selection and oversampling techniques and evaluation of performance with different metrics which account for imbalance. Nonetheless, there exist some limitations in this work which are worthwhile to consider. The most serious problem is the imbalanced distribution among problem classes, in which the most relevant group is only one third of the

majority group, i.e. the control group. More realistic results given this problem were acquired using the SMOTE oversampling technique. However, the over-generalization problem in this context can occur, leading to misclassification of the majority class into the minority class caused by the generation of new synthetic samples, which results in lower performance of the control group. To deal with this problem, extensions of the SMOTE implementation have been proposed to mitigate the impact of this problem in the classification performance. For instance, borderline-SMOTE [Han et al., 2005] gives priority to find the samples which are on the boundaries between classes, i.e. borderline samples, to oversample those in order to achieve better prediction, as frequently the classification algorithms are based on learning these boundaries during training. Another possibility is given by the adaptive synthetic sampling approach (ADASYN) [Haibo He et al., 2008], which according to the distribution of the data, focus on the minority samples which are more difficult to learn. Namely, the number of synthetic samples generated from each minority sample is directly proportional to the number of neighbors of the majority class within a limit distance. In the line of the previous ideas, new versions of these algorithms have been implemented to improve performance, such the density-based SMOTE (DB-SMOTE) [Bunkhumpornpat et al., 2012] or the majority weighted SMOTE (MWMOTE) [Barua et al., 2014]. These methods would allow to have a deeper insight of the distribution of the minority group before and after oversampling the training set.

On the other hand, another limiting factor of this study was the high correlation between features, even though it was accounted for in some degree within the feature selection algorithms. This fact is yet inherent to the problem itself, as it is not suppose to exist a clear difference between retinal layers, less between contiguous macular regions.

Finally, it is important to note the impact of the age on retinal thinning. Although the control group were recruited so as to match the age range of the patients, the effect of this variable requires to be accounted for within the learning algorithm. This way, it would only be assessed the classification ability of retinal thinning linked specially to the disease progression.

In conclusion, although further research is required to validate retinal screening as a diagnositc tool in Parkinson's disease, this work provides factual support to the hypothesis of retinal thinning as part of the clinical picture in Parkinson's Disease.

Appendices

A. APPENDIX

Appendix

A.1 Complementary figures

A.1.1 Box plot outlier detection method

A.1.2 Histrogram distribution plot

A.1.3 Heatmap with correlation values

A.1.4 Univariate filter rank - whole dataset

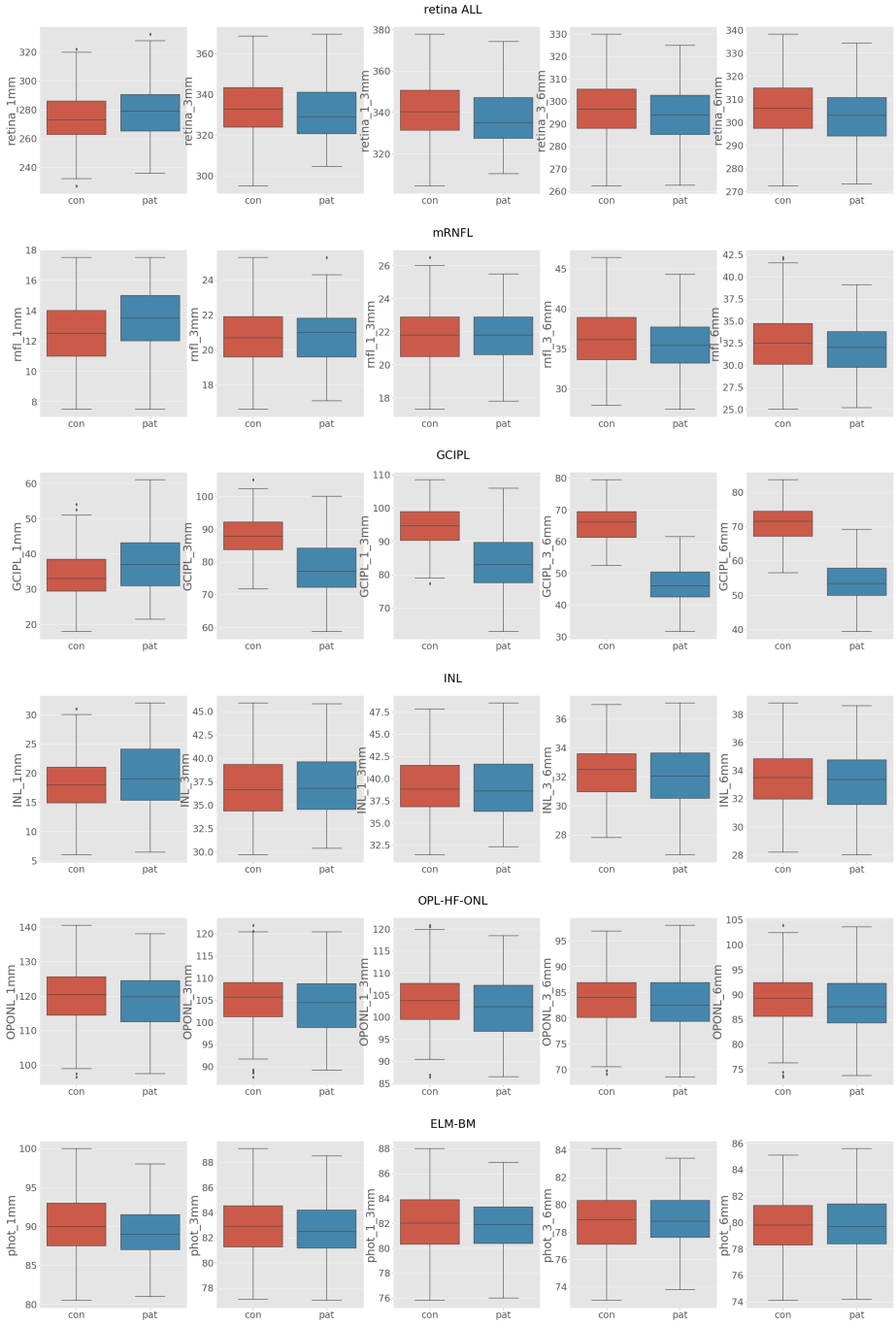


Figure A.1: Box plot diagram of each variable conditioned to each class group.
Retina ALL, total thickness value of the retina; mRNFL, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer; INL, inner nuclear layer; OPL-HF-ONL, outter plexiform-outer nuclear layer; ELM-BM, photoreceptor layer. Thickness within 1-, 3- and 6-mm diameter macular discs and in concentric parafoveal (1-to 3-mm) and perifoveal (3-to 6-mm) rings.

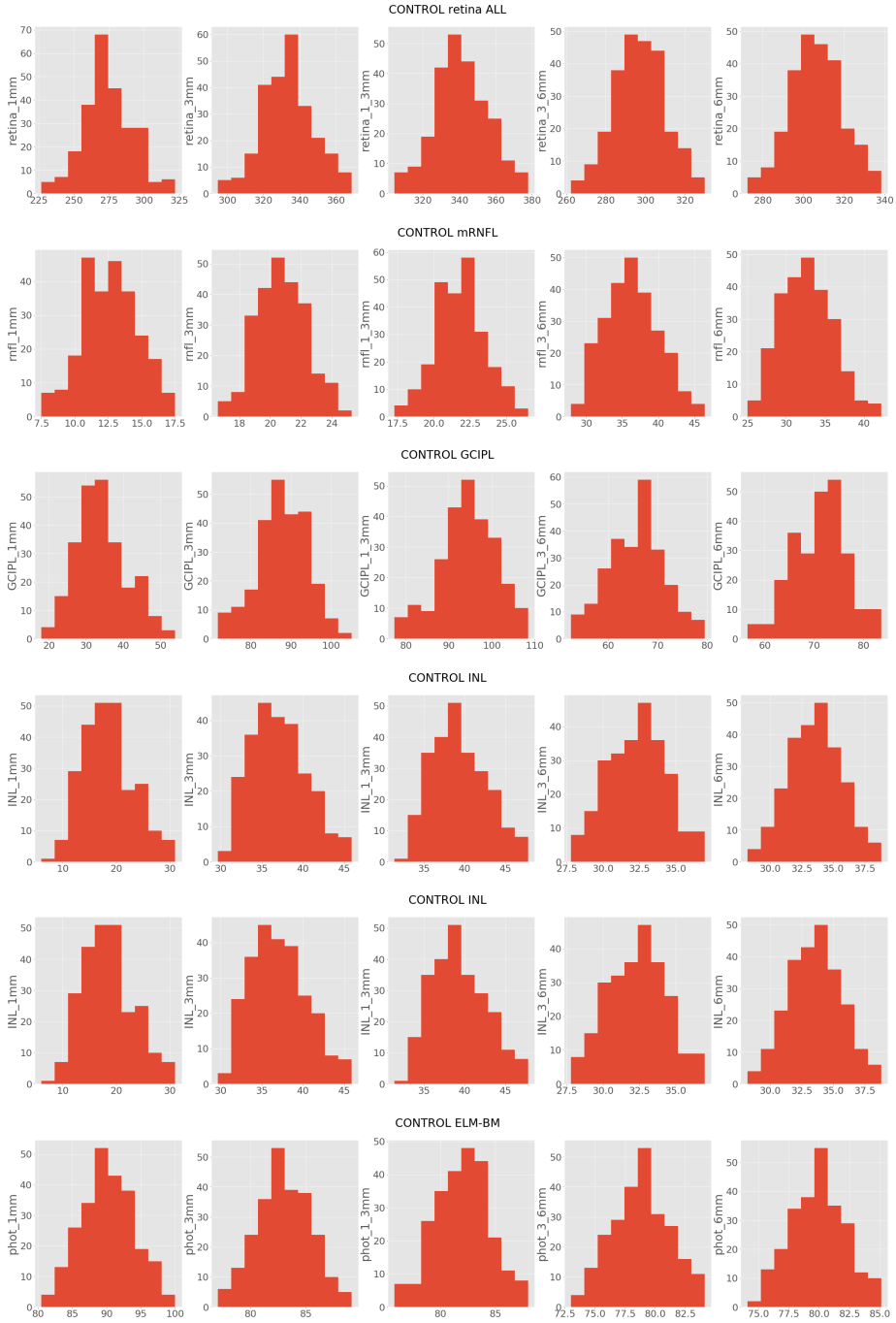


Figure A.2: Histogram distribution of each variable in control data.
Retina ALL, total thickness value of the retina; mRNFL, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer; INL, inner nuclear layer; OPL-HF-ONL, outter plexiform-outer nuclear layer; ELM-BM, photoreceptor layer. Thickness within 1-, 3- and 6-mm diameter macular discs and in concentric parafoveal (1-to 3-mm) and perifoveal (3-to 6-mm).

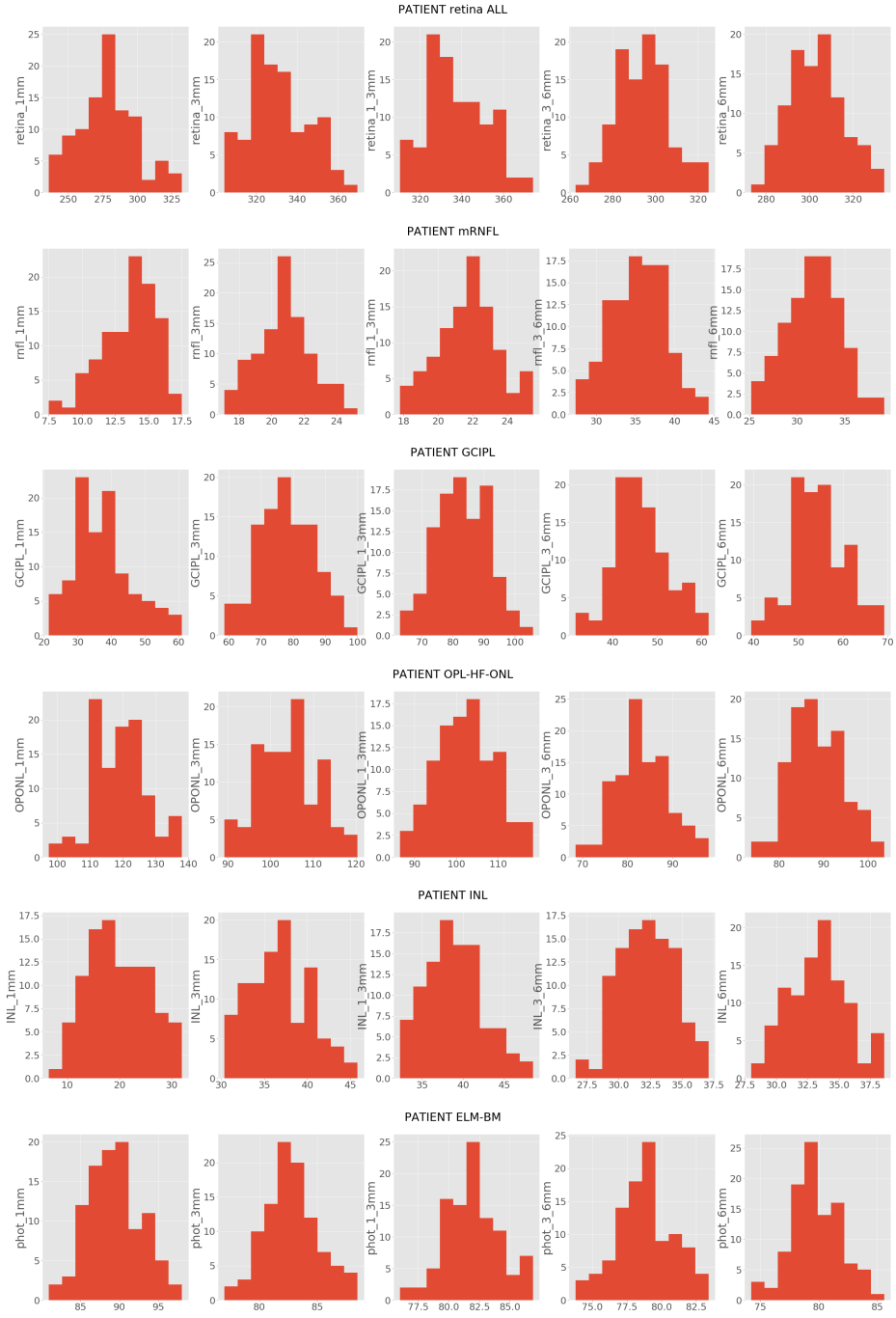
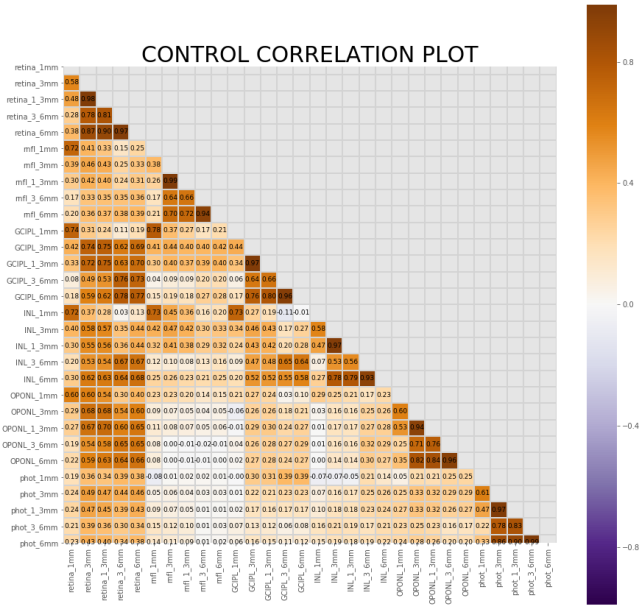
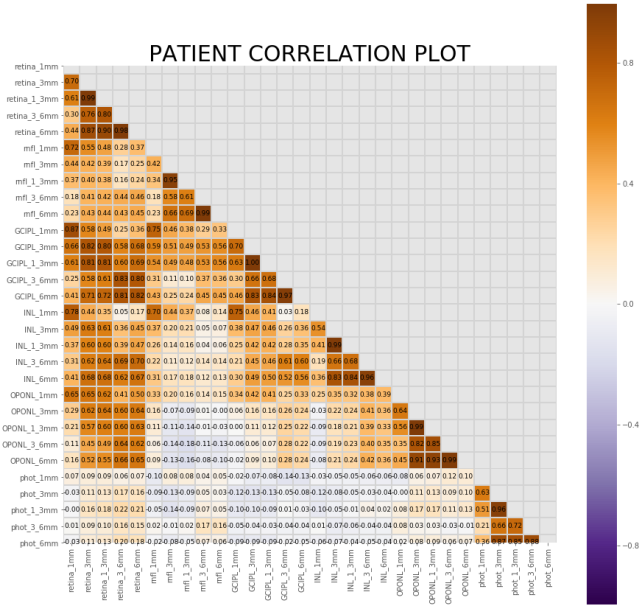


Figure A.3: Histogram distribution of each variable in patient data.
Retina ALL, total thickness value of the retina; mRNFL, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer; INL, inner nuclear layer; OPL-HF-ONL, outter plexiform-outer nuclear layer; ELM-BM, photoreceptor layer. Thickness within 1-, 3- and 6-mm diameter macular discs and in concentric parafoveal (1-to 3-mm) and perifoveal (3-to 6-mm).

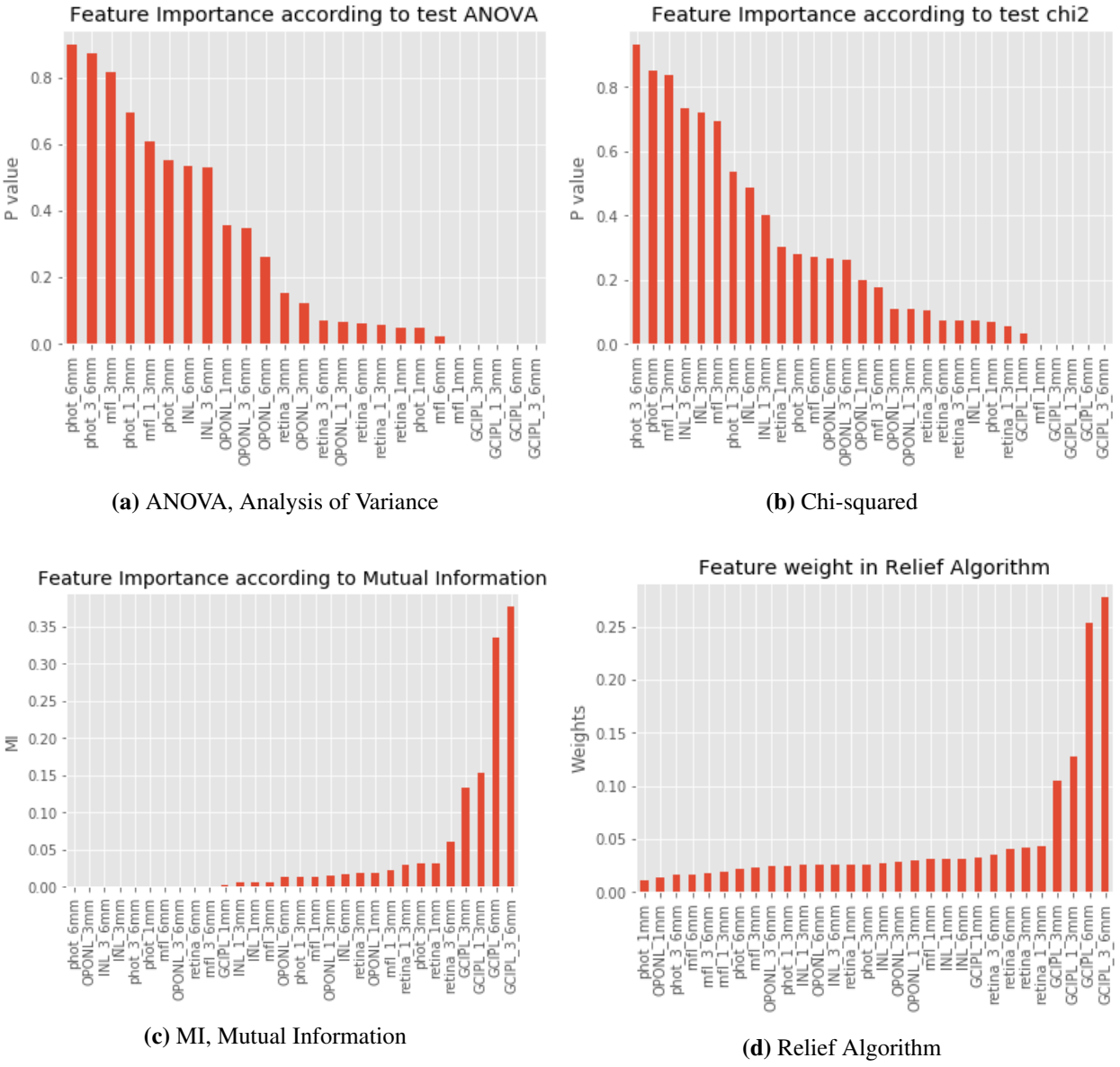


(a) Control data



(b) Patient data

Figure A.4: Heat map with explicit Pearson correlation values in control (a) and patient (b) data. Retina, total thickness value of the retina; mRNFL, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer; INL, inner nuclear layer; OPL-HF-ONL, outter plexiform-outer nuclear layer; ELM-BM, photoreceptor layer; {6-,3-,1-mm}, 6-,3-,1-mm discs; {1_3mm, 3_6mm}, 1-to 3-mm and 3-to 6-mm rings.



(a) ANOVA, Analysis of Variance

(b) Chi-squared

(c) MI, Mutual Information

(d) Relief Algorithm

Figure A.5: Feature selection: univariate filter ranking scores on the whole dataset.
Parametric methods: ANOVA (Analysis of Variance) (a). Non-parametric methods: Chi-Squared (b), MI (Mutual Information) (c) and relief algorithm (d). Retina, total thickness value of the retina; rnfl, retinal nerve fiber layer, GCIPL, ganglion cell-inner plexiform layer; INL, inner nuclear layer; OPONL, outter plexiform-outer nuclear layer; phot, photoreceptor layer; ; KNN, K-Nearest Neighbors Classifier; GaussianNB, Naïve Bayes Classifier; SVM, Support-Vector Machine classifier.

A.2 Complementary tables

A.2.1 D'Angostino K^2 Normality test

Table A.1: Results of D'Agostino and Pearson's omnibus test of normality in **patient** data with **normal** distribution.
Nomenclature according to layer and region and showing the values of the $Z-statistic$ and $p-value$.

Layer	Region (mm)	Z-statistic	P-value
Patient			
retina	1	1.0236	0.5994
	3	2.8660	0.2386
	1 3	3.1446	0.2076
	3 6	0.6349	0.728
	6	0.8642	0.6491
mRNFL	1	5.2543	0.0723
	3	0.0202	0.9899
	1 3	0.1864	0.911
	3 6	0.2027	0.9036
	6	0.2505	0.8823
GCIPL	3	0.957	0.6197
	1 3	0.5393	0.7636
	3 6	1.6189	0.4451
	6	0.7904	0.6736
INL	1	4.4293	0.1092
	3	2.7866	0.2483
	1 3	2.6986	0.2594
	3 6	0.3471	0.8407
	6	1.0013	0.6062
OPONL	1	0.5512	0.7591
	3	1.0406	0.5943
	1 3	1.6374	0.441
	3 6	0.6189	0.7338
	6	0.8204	0.6635
phot	1	0.7553	0.6855
	3	1.0305	0.5974
	1 3	0.0822	0.9597
	3 6	0.0358	0.9822
	6	0.542	0.7626

Table A.2: Results of D'Agostino and Pearson's omnibus test of normality in **control** data with **normal** distribution.
Nomenclature according to layer and region and showing the values of the $Z - statistic$ and $p - value$.

Layer	Region	Z-statistic	P-value
Control			
retina	1	0.4872	0.7837
	3	0.0291	0.9855
	1 3	0.1167	0.9432
	3 6	0.4736	0.7891
	6	0.3054	0.8583
mRNFL	1	1.8350	0.3995
	3	1.2191	0.5436
	1 3	1.4457	0.4853
	6	5.5124	0.0635
GCIPL	3	1.1601	0.5598
	1 3	3.2607	0.1958
	3 6	1.8896	0.3888
	6	2.3237	0.3129
INL	3 6	2.8592	0.2393
	6	2.3554	0.3079
OPONL	1	0.5307	0.7669
	3	0.2380	0.8878
	1 3	0.7646	0.6823
	3 6	0.1428	0.9311
	6	0.0176	0.9912
phot	1	2.5990	0.2726
	3	1.2521	0.5346
	1 3	1.2511	0.5349
	3 6	2.8119	0.2451
	6	2.2331	0.3274

Bibliography

[Andrea, 2013] Andrea, Kliton & Shevlyakov, G. S. P. (2013). Detection of outliers with boxplots. 141-144. pages 141–144.

[Appice et al., 2004] Appice, A., Ceci, M., Rawles, S., and Flach, P. (2004). Redundant feature elimination for multi-class problems. In *Twenty-first international conference on Machine learning - ICML '04*, page 5, Banff, Alberta, Canada. ACM Press.

[Barua et al., 2014] Barua, S., Islam, M. M., Yao, X., and Murase, K. (2014). MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):405–425.

[Bodis-Wollner, 2013] Bodis-Wollner, I. (2013). Foveal vision is impaired in Parkinson's disease. *Parkinsonism & Related Disorders*, 19(1):1–14.

[Bodis-Wollner et al., 2014] Bodis-Wollner, I., Miri, S., and Glazman, S. (2014). Venturing into the no-man's land of the retina in Parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, 29(1):15–22.

[Bodis Wollner and Yahr, 1978] Bodis Wollner, I. and Yahr, M. D. (1978). Measurements of Visual Evoked Potentials in Parkinson's Disease. *Brain*, 101(4):661–671.

[Bunkhumpornpat et al., 2012] Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2012). DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique. *Applied Intelligence*, 36(3):664–684.

[Cao et al., 2014] Cao, P., Yang, J., Li, W., Zhao, D., and Zaiane, O. (2014). Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD. *Computerized Medical Imaging and Graphics*, 38(3):137–150.

[Cawley and Talbot, 2010] Cawley, G. C. and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Chrysou et al., 2019] Chrysou, A., Jansonius, N. M., and van Laar, T. (2019). Retinal layers in Parkinson's disease: A meta-analysis of spectral-domain optical coherence tomography studies. *Parkinsonism & Related Disorders*, 64:40–49.

[Cochran, 1952] Cochran, W. G. (1952). The chi-squared Test of Goodness of Fit. *The Annals of Mathematical Statistics*, 23(3):315–345.

[Cruz-Herranz et al., 2016] Cruz-Herranz, A., Balk, L. J., Oberwahrenbrock, T., Saidha, S., Martinez-Lapiscina, E. H., Lagreze, W. A., Schuman, J. S., Villoslada, P., Calabresi, P., Balcer, L., Petzold, A., Green, A. J., Paul, F., Brandt, A. U., and Albrecht, P. (2016). The APOSTEL recommendations for reporting quantitative optical coherence tomography studies. *Neurology*, 86(24):2303–2309.

[Dawson, 2011] Dawson, R. (2011). How Significant Is A Boxplot Outlier? *Journal of Statistics Education, Volume 19, Number2*.

[Dubis et al., 2012] Dubis, A. M., Hansen, B. R., Cooper, R. F., Beringer, J., Dubra, A., and Carroll, J. (2012). Relationship between the foveal avascular zone and foveal pit morphology. *Investigative Ophthalmology & Visual Science*, 53(3):1628–1636.

[Dudani, 1976] Dudani, S. A. (1976). The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327.

[D'Agostino, 1971] D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large sample size. *Biometrika, 58, 341-348*.

[Fernández et al., 2018] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing, Cham.

[Garcia-Martin E, 2014] Garcia-Martin E, Rodriguez-Mena D, S. M. e. a. (2014). Electrophysiology and optical coherence tomography to evaluate Parkinson disease severity. *Investigative Ophthalmology & Visual Science*, 55(2):696–705.

[Gareth James, 2014] Gareth James, Daniela Witten, T. H. a. R. T. (2014). An Introduction to Statistical Learning: with Applications in R. page 181. Springer Publishing Company, Incorporated.

[Gou et al., 2012] Gou, J., Du, L., Zhang, Y., and Xiong, T. (2012). A New Distance-weighted k-nearest Neighbor Classifier. *Journal of Information and Computational Science*, 9:1429–1436.

[Guan et al., 2009] Guan, D., Yuan, W., Lee, Y.-K., and Lee, S. (2009). Nearest neighbor editing aided by unlabeled data. *Information Sciences*, 179(13):2273–2282.

[Guyon, 2003] Guyon, Isabelle & Elisseeff, A. (2003). An Introduction of Variable and Feature Selection. *Journal of Machine Learning Research*, 1.

[Haibo He et al., 2008] Haibo He, Yang Bai, Garcia, E. A., and Shutao Li (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, Hong Kong, China. IEEE.

[Haindl, 2006] Haindl, Michal & Somol, P. . V. D. . K. C. (2006). Feature Selection Based on Mutual Correlation. In *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4225, pages 569–577. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

[Hajee et al., 2009] Hajee, M. E., March, W. F., Lazzaro, D. R., Wolintz, A. H., Shrier, E. M., Glazman, S., and Bodis-Wollner, I. G. (2009). Inner retinal layer thinning in Parkinson disease. *Archives of Ophthalmology (Chicago, Ill.: 1960)*, 127(6):737–741.

[Hall, 1999] Hall, M.A. and Smith, L. (1999). Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper. *American Association of Artificial Intelligence*.

[Han et al., 2005] Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In Huang, D.-S., Zhang, X.-P., and Huang, G.-B., editors, *Advances in Intelligent Computing*, volume 3644, pages 878–887. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

[Hand, 2012] Hand, D. J. (2012). Assessing the Performance of Classification Methods: *Assessing the Performance of Classification Methods*. *International Statistical Review*, 80(3):400–414.

[He and Ma, 2013] He, H. and Ma, Y., editors (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley, 1 edition.

[Hodges and Lehmann, 1963] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics*, 34(2):598–611.

[Hoehn and Yahr, 1967] Hoehn, M. M. and Yahr, M. D. (1967). Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427.

[Jiménez et al., 2014] Jiménez, B., Ascaso, F. J., Cristóbal, J. A., and López del Val, J. (2014). Development of a prediction formula of Parkinson disease severity by optical coherence tomography. *Movement Disorders: Official Journal of the Movement Disorder Society*, 29(1):68–74.

[Jovic et al., 2015] Jovic, A., Brkic, K., and Bogunovic, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, Opatija, Croatia. IEEE.

[Kao and Green, 2008] Kao, L. S. and Green, C. E. (2008). Analysis of Variance: Is There a Difference in Means and What Does It Mean? *Journal of Surgical Research*, 144(1):158–170.

[Knighton et al., 2012] Knighton, R. W., Gregori, G., and Budenz, D. L. (2012). Variance Reduction in a Dataset of Normal Macular Ganglion Cell Plus Inner Plexiform Layer Thickness Maps with Application to Glaucoma Diagnosis. *Investigative Opthalmology & Visual Science*, 53(7):3653.

[Kourou et al., 2015] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

[Kuncheva, 2017] Kuncheva, L. I. (2017). A STABILITY INDEX FOR FEATURE SELECTION. page 6.

[Kuncheva and Rodríguez, 2018] Kuncheva, L. I. and Rodríguez, J. J. (2018). On feature selection protocols for very low-sample-size data. *Pattern Recognition*, 81:660–673.

[Lang et al., 2013] Lang, A., Carass, A., Hauser, M., Sotirchos, E. S., Calabresi, P. A., Ying, H. S., and Prince, J. L. (2013). Retinal layer segmentation of macular OCT images using boundary classification. *Biomedical Optics Express*, 4(7):1133.

[Lavrač et al., 1999] Lavrač, N., Gamberger, D., and Jovanoski, V. (1999). A study of relevance for learning in deductive databases. *The Journal of Logic Programming*, 40(2-3):215–249.

[Lehmann, 2013] Lehmann, R. (2013). 3-Rule for Outlier Detection from the Viewpoint of Geodetic Adjustment. *Journal of Surveying Engineering*, 139(4):157–165.

[Li et al., 2017] Li, S.-t., Wang, X.-n., Du, X.-h., and Wu, Q. (2017). Comparison of spectral-domain optical coherence tomography for intra-retinal layers thickness measurements between healthy and diabetic eyes among Chinese adults. *PLOS ONE*, 12(5):e0177515.

[Markowski and Markowski, 1990] Markowski, C. A. and Markowski, E. P. (1990). Conditions for the Effectiveness of a Preliminary Test of Variance. *The American Statistician*, 44(4):322.

[Medical Advisory Secretariat, 2009] Medical Advisory Secretariat (2009). Optical coherence tomography for age-related macular degeneration and diabetic macular edema: an evidence-based analysis. *Ontario Health Technology Assessment Series*, 9(13):1–22.

[Murueta-Goyena et al., 2019] Murueta-Goyena, A., Del Pino, R., Reyero, P., Galdós, M., Arana, B., Lucas-Jiménez, O., Acera, M., Tijero, B., Ibarretxe-Bilbao, N., Ojeda, N., Peña, J., Cortés, J., Gómez-Esteban, J. C., and Gabilondo, I. (2019). Parafoveal thinning of inner retina is associated with visual dysfunction in Lewy body diseases. *Movement Disorders: Official Journal of the Movement Disorder Society*, 34(9):1315–1324.

[Pelc et al., 2019] Pelc, M., Khoma, Y., and Khoma, V. (2019). ECG Signal as Robust and Reliable Biometric Marker: Datasets and Algorithms Comparison. *Sensors*, 19(10):2350.

[Press, 2007] Press, W. H., editor (2007). *Numerical recipes: the art of scientific computing*. Cambridge University Press, Cambridge, UK ; New York, 3rd ed edition. OCLC: ocn123285342.

[Provis et al., 2005] Provis, J. M., Penfold, P. L., Cornish, E. E., Sandercoe, T. M., and Madigan, M. C. (2005). Anatomy and development of the macula: specialisation and the vulnerability to macular degeneration. *Clinical and Experimental Optometry*, 88(5):269–281.

[Puntumapon and Waiyamai, 2012] Puntumapon, K. and Waiyamai, K. (2012). A Pruning-Based Approach for Searching Precise and Generalized Region for Synthetic Minority Over-Sampling. In *Advances in Knowledge Discovery and Data Mining*, volume 7302, pages 371–382. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

[Raschka, 2018] Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software*, 3(24):638.

[Ross, 2014] Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS One*, 9(2):e87357.

[Saeys et al., 2007] Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[Santafe et al., 2015] Santafe, G., Inza, I., and Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508.

[Shalev-Shwartz, 2014] Shalev-Shwartz, Shai, B.-D. S. (2014). 18. Decision Trees. *Understanding Machine Learning. Cambridge University Press*.

[Shannon, 1948] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

[Springer AD, 2005] Springer AD, H. A. (2005). Development of the primate area of high acuity. pages 1. Use of finite–element analysis models to identify mechanical variables affecting pit formation(21:53–62), 2. Quantitative morphological changes associated with retina and pars plana growth(21:775–790), 3. Temporal relationships between pit formation, retinal elongation and cone packing(22:171–185) . Vis Neurosci.

[Strimbu and Tavel, 2010] Strimbu, K. and Tavel, J. A. (2010). What are biomarkers?:. *Current Opinion in HIV and AIDS*, 5(6):463–466.

[Tewarie et al., 2012] Tewarie, P., Balk, L., Costello, F., Green, A., Martin, R., Schippling, S., and Petzold, A. (2012). The OSCAR-IB Consensus Criteria for Retinal OCT Quality Assessment. *PLoS ONE*, 7(4):e34823.

[Tian et al., 2011] Tian, T., Zhu, X.-H., and Liu, Y.-H. (2011). Potential role of retina as a biomarker for progression of Parkinson's disease. *International Journal of Ophthalmology*, 4(4):433–438.

[Tukey, 1977] Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co, Reading, Mass.

[Urbanowicz et al., 2018] Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., and Moore, J. H. (2018). Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining. *arXiv:1711.08477 [cs]*. arXiv: 1711.08477.

[Zarranz JJ, 2004] Zarranz JJ, Alegre J, G.-E. J. e. a. (2004). The new mutation, E46K, of -synuclein causes parkinson and Lewy body dementia: New -Synuclein Gene Mutation. *Annals of Neurology*, 55(2):164–173.

[Zhang, 2004] Zhang, H. (2004). The Optimality of Naive Bayes. In V. Barr & Z. Markov (eds.). *Proceedings of the Seventeenth International Florida Artificial Intelligence research Society Conference (FLAIRS 2004), : AAAI Press*.

[Zhang et al., 2013] Zhang, Y., Li, S., Wang, T., and Zhang, Z. (2013). Divergence-based feature selection for separate classes. *Neurocomputing*, 101:32–42.

[Zhao and Morstatter, 2010] Zhao, Z. and Morstatter, F. (2010). Advancing Feature Selection Research. *ASU Feature Selection Repository Arizona State University*, pages 1–28.

[Z.S. Nasreddine, 2005] Z.S. Nasreddine, N.A. Phillips, V. B. S. C. V. W. I. C. e. a. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment: MOCA: A BRIEF SCREENING TOOL FOR MCI. *Journal of the American Geriatrics Society*, 53(4):695–699.