UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Master in Innovation and Research in Informatics
Data Science

Using machine learning on the sources of retinal images for diagnosis by proxy of diabetes mellitus and diabetic retinopathy

---

Master Thesis

Author: Anass Benali Bendahmane

Advisor: Alfredo Vellido Alcacena

June 2021

Contents

Glossary

black box model a method which learns to map inputs to outputs, where the learned mapping is not readily interpretable.

boxplot A method for graphically depicting groups of numerical data through their quartiles. It allows to get insights of how the numerical data values are spread.

components Learnt features from a matrix decomposition method. In other descriptions they are also known as sources or basis.

convex optimization problem An optimization problem in which the objective function is a convex function and the feasible set is a convex set.

cost function Function used to determine the error (loss or distance) between an output and a target. It is also known as loss function or error function.

frobenius norm An extension of the Euclidean norm to matrices.

image mask a filter which determines the region of interest in the image.

rgb color model an additive color model in which red, green, and blue light are added together in different proportions to reproduce a broad variety of colours.

supervised learning the task of learning patterns from labeled data.

unsupervised learning the task of learning patterns from unlabeled data.

Acronyms

AUC Area Under Curve.

CV Cross-Validation.

DL Deep Learning.

FAZ Foveal Avascular Zone.

ICA Independent Component Analysis.

LDA Linear Discriminant Analysis.

LDR Linear Dimensionality Reduction.

LR Logistic Regression.

MI Mutual Information.

ML Machine Learning.

MUR Multiplicative Update Rules.

NMF Non-negative matrix factorization.

NNDSVD Nonnegative Double Singular Value Decomposition.

PCA Principal Component Analysis.

RBF Radial basis function.

ROI Region Of Interest.

SSI Signal Strength Index.

SVD Singular Value Decomposition.

SVM Support vector machine.

VQ Vector Quantization.

Abstract

In current research in ophthalmology, images of the vascular system in the human retina are used as exploratory proxies for pathologies affecting different organs. This thesis addresses the analysis, using machine learning and computer vision techniques, of retinal images acquired with different techniques (Fundus retinographies, optical coherence tomography and optical coherence tomography angiography), with the objective of using them to assist diagnostic decision making in diabetes mellitus and diabetic retinopathy. This thesis explores the use of matrix factorization-based source extraction techniques, as the basis to transform the retinal images for classification. The proposed approach consists on preprocessing the images to enable the learning of an unsupervised parts-based representation prior to the classification. As a result of the use of interpretable models, with this approach we unveiled an important bias in the data. After correcting for the bias, promising results were still obtained which merit for further exploration.

Graphical abstract

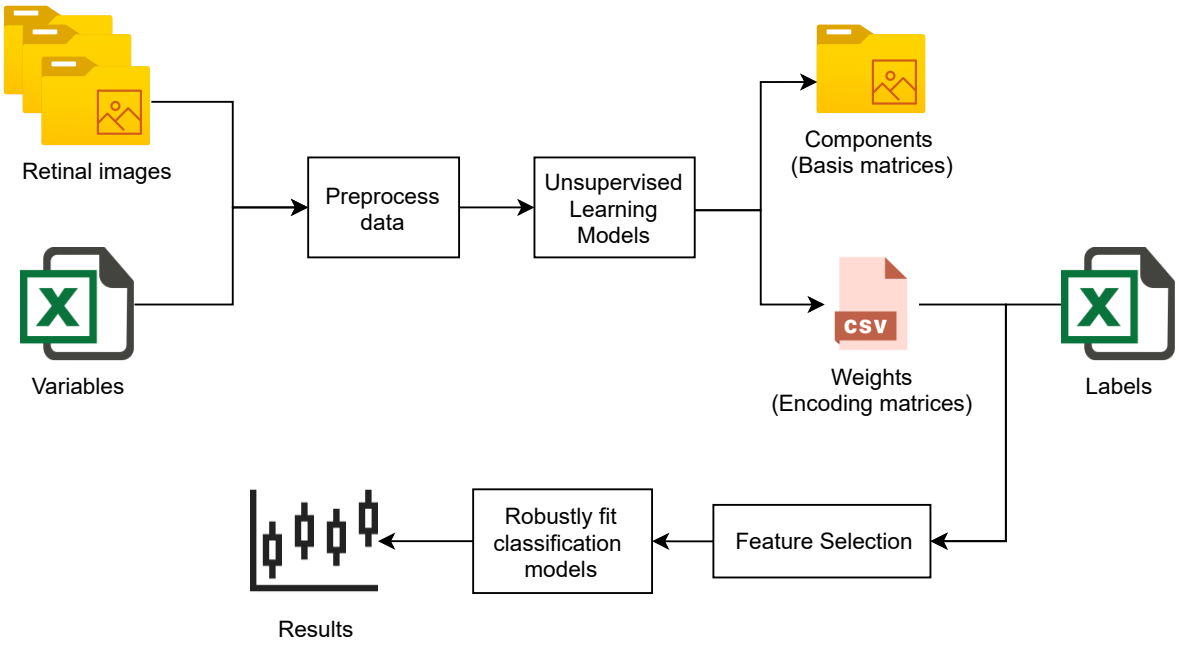


Figure 1: Graphical outline of the data analysis workflow of the experiments reported in the current thesis.

Chapter 1

Introduction

Diabetic Retinopathy (DR) is the leading cause of human blindness in Type 1 Diabetes Mellitus (DM) patients, as a consequence of impaired blood flow in the retina [1].

Type 1 DM is a serious and lifelong condition. Neither the cause of the condition nor the means to prevent it are known [2]. Although it can appear at any age, it typically begins in children and young adults [3]. Type 1 DM is estimated to constitute between 5 to 10% of all diabetes cases [4].

DM is a metabolic disorder that causes high levels of sugar in blood, also known as Hyperglycemia. DM is usually diagnosed by testing the level of sugar (HbA1C) in the blood [5]. There are several different types of diabetes, one of them being Type 1 DM. Type 1 DM is characterized by the pancreas inability to generate enough insulin, a hormone required by the human body cells to use sugar as a source of energy. This condition can lead to complications and, when those affect the blood vessels in the retina, it can develop in what its known as DR, which may cause several vision difficulties [6].

In its early stages, DR may cause no symptoms, or, at most, only mild vision problems, which makes it hard to detect. Even if symptoms reveal themselves, and the patient is subsequently referred to a doctor, DR is still hard to diagnose, as the differences between a healthy eye and an eye with early-stage DR are rather small. Detecting DR early on after onset is thus very important in order to slow its advancement, or even prevent the vision complications which can lead to blindness if left untreated.

Different non-invasive imaging techniques can be applied to the study of retinal diseases in general, and DR in particular, such as fundus retinography (FR, Figure 1.1), structural Optical Coherence Tomography (OCT, Figure 1.2), or Optical Coherence Tomography Angiography (OCTA, Figure 1.3).

Figure 1.1: Fundus Retinography

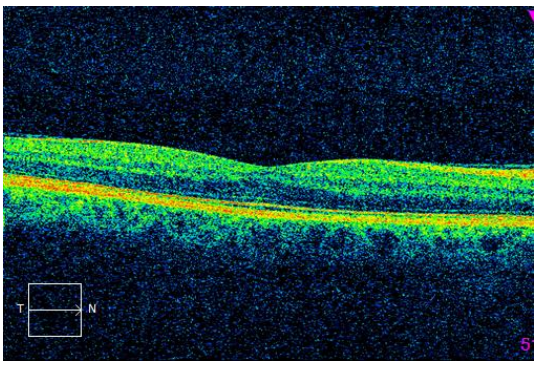


Figure 1.2: Optical Coherence Tomography

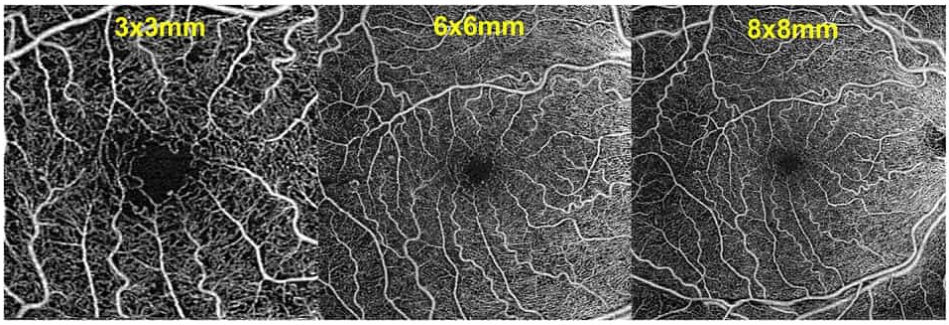


Figure 1.3: Optical Coherence Tomography Angiography

Traditionally, standard DR screening systems use retinography [7] due to its widespread availability. For this reason, the vast majority of computational science applications in ophthalmology have been applied to FR, and far more rarely to OCT scans, which are not so readily available. Recently, the advent of the more advanced OCTA technology (still available only in a limited number of clinical outlets) allows for direct visualization of flow in the retinal vessels, representing a significant advance in the evaluation of these patients.

This thesis aims to analyze a high-quality multi-modal image dataset gathered throughout previous ophthalmology research projects (*Fundació La Marató TV3, Fondo Investigaciones Sanitarias, FIS*). This work substantiates a collaboration with Dr. Javier Zarranz-Ventura (Institut Clinic d'Oftalmologia, ICOF, Hospital Clinic de Barcelona), who has provided the annotated database that is the basis of the thesis.

Some recent and ongoing work with the aforementioned dataset suggests that a feature extraction approach based on radiomics [8] shows reasonably good discriminative

capabilities for DR. Still, the interpretability of the results is important in the medical field, and so in this thesis we explore the use of a source extraction approach for data transformation, focusing on linear dimensionality reduction (LDR) techniques to extract visually interpretable results. This way, by taking a different feature extraction approach we aim to find out whether the pathology discrimination results based on radiomics are reproducible and maybe even if they can be improved.

Among the many matrix factorization techniques based on LDR, this work mostly focuses in non-negative matrix factorization (NMF) and some of its variants. NMF has found great success in learning interpretable parts-based representations in areas such as facial recognition, recommenders, or astronomy, to name a few.

1.1 Objectives

This thesis aims to analyze the potential of source extraction techniques, mainly by looking at NMF and some of its variants, as a feature extraction approach for the available retinal image dataset. The adequacy of the approach will be evaluated by the performance of classifying different levels of DR. Specifically, the evaluation will be based on the discrimination ability of the learnt features for two relevant tasks: To discriminate between non-diabetics (controls) and diabetics and to discriminate between the absence or presence of DR. To achieve these aims, we define the following set of objectives:

- Review the concepts for understanding the imaging techniques and the literature of application of non-negative source extraction.
- Explore the dataset to analyze and fix any incoherence and nuances found. This is important so that that future work can be more focused on the data modeling itself, from a classification perspective.
- Develop the necessary preprocessing and image transformations for each type of retinal image so that the unsupervised factorization models can learn useful representations.
- Develop a robust feature extraction scheme to make use of the three types of retinal images in the classification stage.
- Obtain insights about the retinal images to try to understand and compare the results with a recent radiomics feature extraction approach.

- Assess how well suited are source extraction based on LDR techniques and especially NMF for usage on a retinal image dataset, based on the results.

1.2 State of the art

Most of the existing recent related work applying computational intelligence methods to DR does not make distinction of the underling diabetes type. According to the expertise of the ophtalmologist advising the thesis, detecting DR for Type 1 Diabetes is a far harder problem than that concerning the other types of diabetes. As previously stated, most of the work in the field involves investigating FR images or clinical data, mainly from Type 2 DM patients datasets, as this is the most frequent type of DM (up to 90% of cases). In this section, we review some of the related work so as to provide a viewpoint of how retinal imaging and source extraction (based on matrix factorization) is usually employed.

Fenner *et al.* provide a complete review of advances in retinal imaging [9]. It goes through how recent studies used FR, OCT and OCTA imaging for DR and other retinal complications. An interesting use case of OCTA images is that concerning the delineation of the foveal avascular zone (FAZ) (Figure 1.4) and how its acircularity relates to DR [10]. Also the FAZ characteristics can be used for quantitative assessment and monitoring of DR [11].

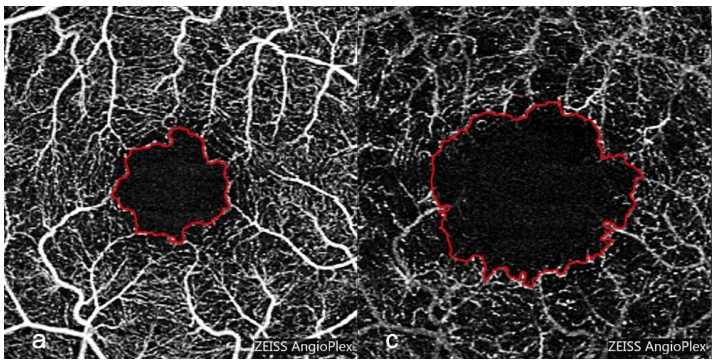


Figure 1.4: Evaluation of the foveal avascular zone (FAZ) [12]

The review also points to an interesting problem with the reproducibility and robustness of black-box models such as the Google Brain system [13]. It does so by citing the work of Lynch *et al.* [14] which shows how Deep Learning (DL) approaches are vulnerable to seemingly innocuous alterations of fundus images. It summarizes the aforementioned problem as follows:

Using a series of reference DR fundus photos containing typical features of DR, slight pixel modifications (0.12–0.51% of total pixels) were introduced into the images that were essentially imperceptible to human readers and did not obviously alter the appearance of the DR lesions. [...] The modified images were classified as normal by the image-based CNN while the lesion-based system still detected DR.

Finally, Fenner *et al.* explain the prohibitive costs of expensive camera equipment, specially for OCT and OCTA.

The monography [15] provides a summary of the current state of the art of applied artificial intelligence (AI) systems in ophthalmology. It reviews many of the AI models and how they are trained and applied to different types of retinal imaging modalities and different retinal diseases, including DR.

The authors of [16] make use of a 3D OCT tensor, using a novel DL framework with two stages: first, a segmentation network; and second, a classification network. The authors claim this framework to tackle the generalization to a new scanning device thanks to having the second stage being device-independent. Furthermore, the study claims that the framework has an error rate which outperforms clinical experts in an OCT-only setting when predicting referral urgency of DR.

Two Kaggle competitions for DR diagnosis using FR images have been organized. The first one back on 2015 was won by Ben Graham [17], while the second was organized in 2019 [18] and won by Guanshuo Xu. One interesting observation related to the competition scores is that, back in 2015, the top score was 0.84957 while the top 100 was 0.45082. Instead, in 2019 the top score was 0.936129, while the top 100 score was a close 0.922081. Furthermore, while back in 2015 all the winners preprocessed in some way the images, the 2019 competition winner did not [19]. Guanshuo Xu achieved his solution by carefully training and ensembling eight DL models (Inception and ResNet). He reported that, for image input sizes above 384 pixels, he did not see any significant improvement in results.

In [20], authors perform retinal vessel segmentation with NMF and a 3D U-Net on FR images. They develop a new within-class and between-class constrained NMF algorithm to extract neighborhood feature information of every pixel and reduce feature

data. They discuss their image preprocessing, which consists on extracting the green channel, applying a Gaussian filter, a gamma correction and, finally, region processing.

In another medical (but unrelated) area, the authors of [21] recently applied blind source separation on magnetic resonance spectroscopic imaging (MRSI) data of brain tumors with Convex-NMF. They extracted signal sources that are interpretable given that they are formed as a linear combination of the original data samples. Those source weights were used as features in classification models which perform great for separating healthy from necrotic tissue.

1.3 Structure of the document

The reminder of the thesis document is organized as follows. On Chapter 2 the relationship between the anatomy of the eye and the imaging techniques is explained along with how the retinal images are taken and where they come from. In Chapter 3 the NMF method is presented along with other related factorization methods. On Chapter 4, we summarily describe some computer vision and machine learning concepts that are used in the thesis. In Chapter 5 we explain all the experimental setting of the thesis: The methodology, the description of the dataset, an exploration of how the unsupervised models work, the developed preprocessing steps and, finally, we report the learnt sources and the classification results. In Chapter 6, we gather the conclusions and outline future work.

Chapter 2

Retinal Images

In this chapter, we provide readers with a self-contained summary of the anatomy of the eye, a basic description of diabetic retinography and explanation about the different retinal imaging techniques. The three non-invasive retinal imaging techniques considered in this thesis are FR, OCT and OCTA. They capture the posterior part of the eye, also known as the *fundus*. Each imaging technique achieves a distinct result which comes with its own advantages and disadvantages.

2.1 Anatomy of the eye

In order to understand what each type of image captures, a reminder of the eye general anatomy is bound to be helpful, and is provided in Figure 2.1.

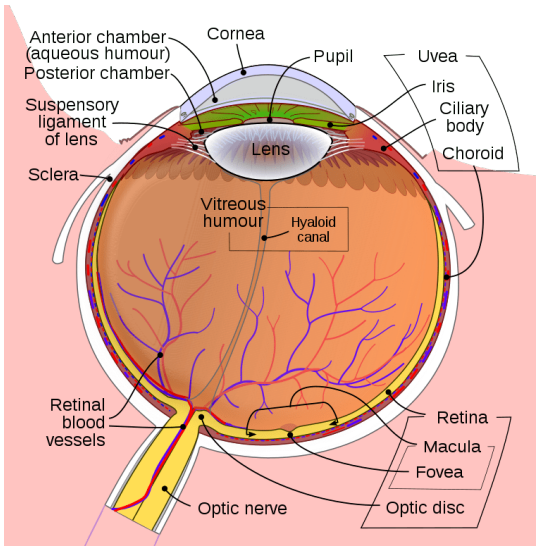


Figure 2.1: Anatomy of the eye [22]

The relevant parts for understanding the imaging techniques are: the fovea, the retina and the choroid. Other parts of interest include the area around the fovea which is called macula and the optic disc which is where the optic nerve exits the retina.

2.2 Diabetic Retinopathy

When an eye is affected with DR, damages will start to develop in the form of hemorrhages, hard exudates, aneurysms, etc. In Figure 2.2, we display a representative example. The dark spot on the left hand side is the macula.

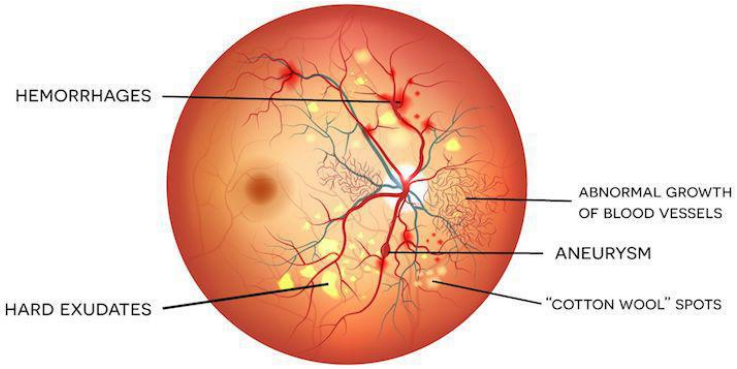


Figure 2.2: DR lesions in the eye [23]

Usually, a clinician rates the presence of DR in each image on a scale of 0 to 4, according to the following description in Table 2.1:

0	No DR
1	Mild
2	Moderate
3	Severe
4	Proliferative DR

Table 2.1: DR rating scale [24]

An eye is categorized as being in one of those levels of severity according to guidelines of the visible lesions on the eye retinal images [25]. Despite that, one concern is that the labeling can have a large inconsistency and, as a consequence, the results obtained for a given dataset will not necessarily generalize to another. This same concern was brought out by the authors of [13] in a presentation at the TensorFlow Dev Summit

2017, where they showed the degree of inconsistency in the labeling of DR. A slide from the presentation can be seen in Figure 2.3. The authors claim to approach the problem by using the consensus over several ratings of different experts, but did not make their dataset public. A later replication study tried the proposed model on public datasets but they were not able to reproduce the results [26]. Therefore, one must be careful, given that the nature of the problem is inherently noisy.

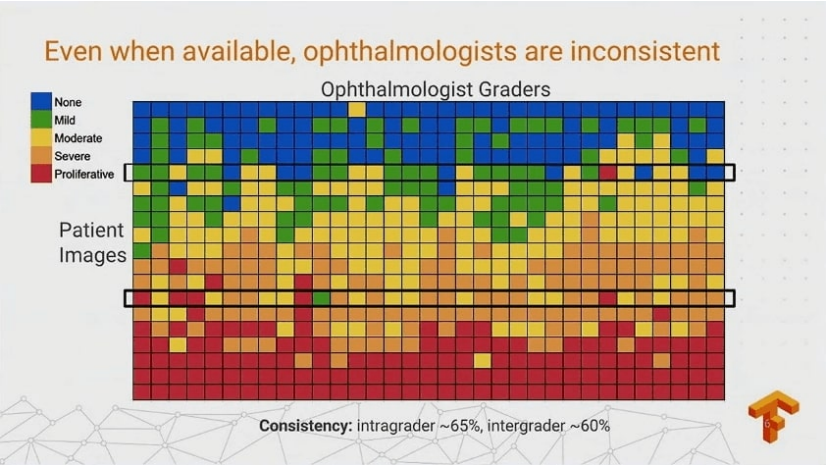


Figure 2.3: Inconsistency of ophthalmologist's estimation [13]

2.3 Fundus Retinography

Fundus Retinography (FR) is a photograph of the posterior part of the eye (the back of the eye). Specialized camera equipment technology produces and digitalizes the image. Figure 2.4 shows how a FR looks like, including some of the eye anatomical features.

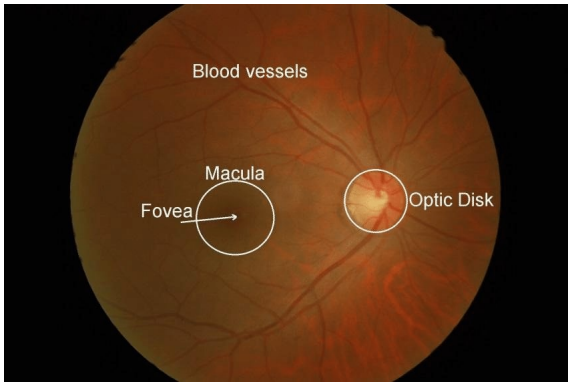


Figure 2.4: FR with marked features [27]

For a right eye, the optic disk will be on the right, whereas for the left eye the

situation is flipped around the vertical axis, being the optic nerve on the left. Therefore, the eye shown in Figure 2.4 is a right eye. Furthermore, unless the FR is centered around the macula, the right eye will have the macula slightly to the left, whereas the left eye will have the macula slightly to the right.

Moreover, the FR can be taken using different fields of view, which can appear as a cropped circle and with different level of zoom. We can see such examples in Figure 2.5.

Figure 2.5: Field of view of FR [28]

The lesions associated with DR will be relatively apparent in the FR images, especially for more advanced stages of DR. An annotated example is provided in Figure 2.6, where the lesions can be appreciated.

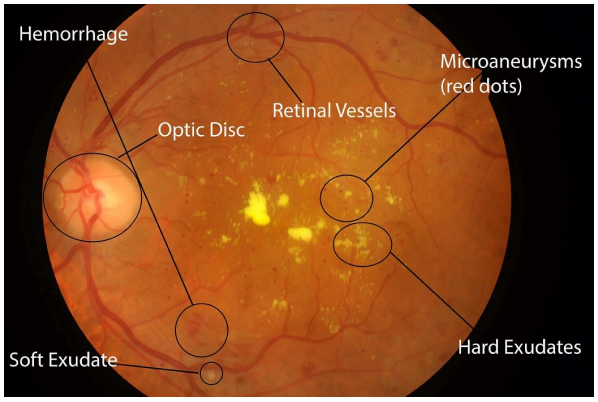


Figure 2.6: DR eye lesions in a FR image [29]

2.4 Optical Coherence Tomography

Optical Coherence Tomography (OCT) uses a type of imaging that produces a three-dimensional cross-section of the eye, allowing to see in the depth of the tissue. Figure 2.7 shows a marked OCT. The center target of the OCT scans for retinal images is usually the fovea.

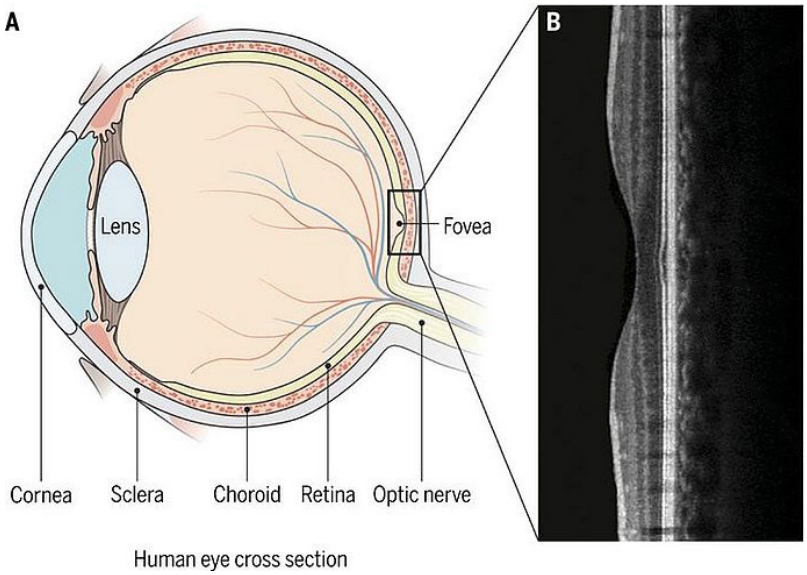


Figure 2.7: OCT scan shows the fundus cross-section around the fovea [30]

OCT relies on tomography, the process of imaging by sections through the use of a penetrating wave [31]. OCT builds a tomogram by using the principles of interferometry and coherent light. By scanning light across the target at each iteration, a depth profile is produced from the time delay and intensity of the backscattered light [32]. This is called an amplitude scan, or A-scan for short. A cross-sectional tomogram can be assembled by laterally combining neighboring A-scans [33]. The assembled tomogram is known as B-scan or brightness scan. Figure 2.8 shows the relative orientation of the A-scan and B-scan.

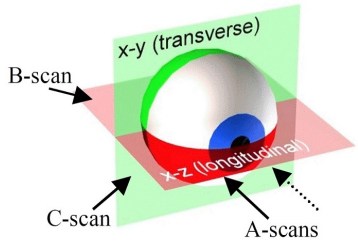


Figure 2.8: Relative orientation of the axial scan (A-scan), longitudinal slice (B-scan) and transversal slice (C-scan) [34]

To sum up, OCT is built by beaming coherent light to obtain A-scans (each of which samples a single slice) and then many neighboring A-scans are assembled to build a two-dimensional B-scan. If we then obtain many neighboring B-scans we would obtain a three-dimensional tomograph [32]. Figure 2.9 graphically depicts the procedure of how the scans are obtained.

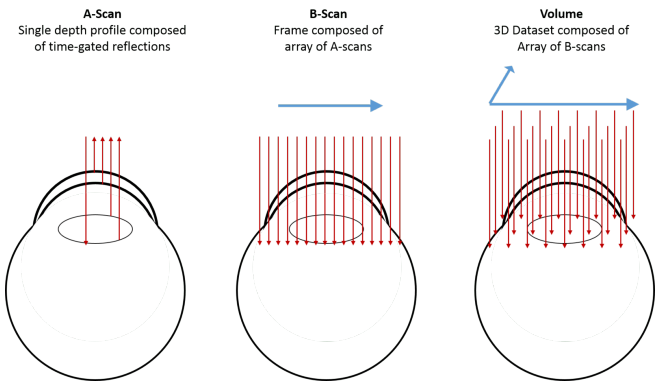


Figure 2.9: Scan types with Optical Coherence Tomography [32]

2.5 Optical Coherence Tomography Angiography

The Optical Coherence Tomography Angiography (OCTA) images are taken by sampling several OCT scans on the same slice many times to estimate the retinal flow. This procedure is repeated for several slices around the fovea.

For the OCTA, the side dimensions of the captured area are usually taken to be either $3 \times 3mm$, $6 \times 6mm$, or $8 \times 8mm$. The OCTA images are then reconstructed in two modalities of image based on the two main vascular plexus of the retina, the *superficial* and *deep* vascular plexus. The *superficial* images allow a good view of the superficial vessels while the *deep* images allow the visualization of the deep vascular plexus, that often is affected in early phases of DR showing areas of impaired blood flow. Moreover, the OCTA images allow to delineate the foveal avascular zone (FAZ) area which is the center dark spot on the OCTA images. On Figure 2.10 we can see an example of progression of DR on OCTA images.

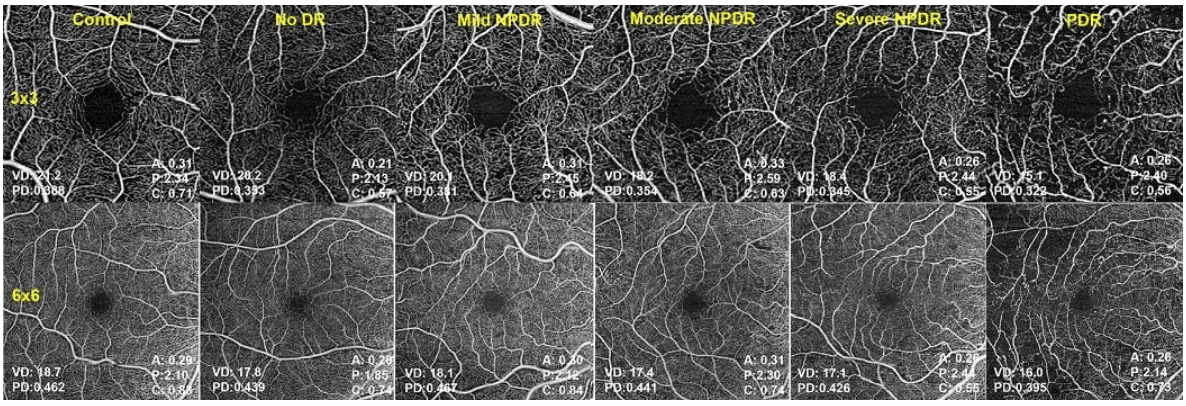


Figure 2.10: DR progression seen in OCTA images [35]

2.6 Relationship between FR, OCT and OCTA

To further understand how the imaging techniques are related, in Figure 2.11 an illustration is provided. The FR image corresponds to the "orange" circle. The OCT image corresponds to a cross-sectional slice centered around the fovea (the darker spot in the center) in the depth direction (capturing some layers of tissue). Finally, the OCTA captures the area of the drawn green square centered in the fovea; it can be seen as a way of capturing the vessels around that area with higher resolution than FR, allowing the estimation of the blood flow and the FAZ area.

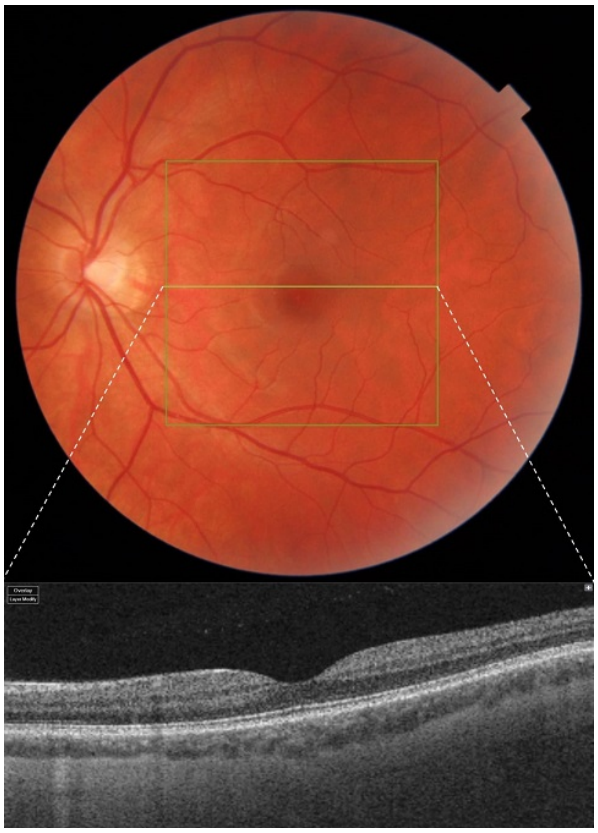


Figure 2.11: Relationship between the three imaging techniques: FR, OCT, OCTA [36]

Chapter 3

Non-negative matrix factorization

Non-negative matrix factorization (NMF) is a widely used blind source separation technique for the analysis of multivariate data. When the conditions are right, it automatically learns a meaningful parts-based representation [37]. Even if the conditions are not quite right, a carefully constrained NMF can produce parts-based representations, resulting in interpretable models [38]. NMF seeks to take advantage of the property that some data is often non-negative by nature, like pixel intensities, user scores, or amplitude spectra, among others. Intuitively, a part-based representation is the notion of combining parts to form a whole [39] and from a signal processing viewpoint, it assumes that the observed data are the manifiestation of a combination of signal sources.

3.1 History

Non-negative matrix factorization was first introduced in 1994 by Paatero and Tapper [40]. They were the first to show interest in the problem and propose approaches to solve it. Later it was popularized by Lee and Seung on 1999 through a paper published in *Nature* for "learning the parts of objects" [39]. Since then, it turned into a widely used tool in various research areas for diverse applications.

Lee and Seung presented an algorithm to solve NMF using multiplicative updates and presented two use cases, learning features on a face and topics from documents. They showed that a decomposition with non-negativity constraints enables a sparse and part-based decomposition to be learnt because cancellation of features is not allowed, only addition. Figure 3.1 shows the original example that Lee and Seung provided in their paper. It can be seen that NMF reconstructs a face from a non-negative basis matrix and encoding matrix.

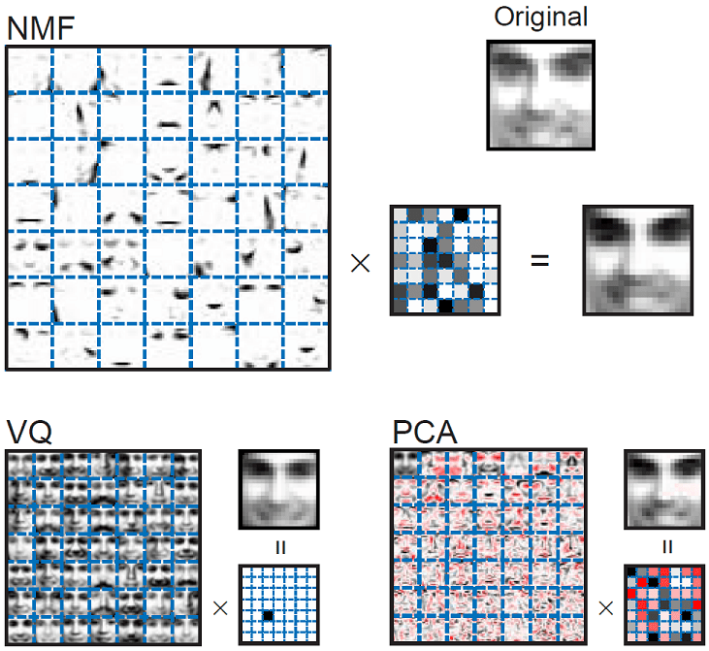


Figure 3.1: Example from Lee and Seung paper which shows how NMF learns a parts-based representation of faces, while vector quantization (VQ) and principal component analysis (PCA) learn *holistic* representations [39].

In essence, NMF can be interpreted as a process of reconstructing multivariate data as a linear combination of the learnt features (sources). That makes the resulting decomposition of the data interpretable as a weighted combination of the features extracted.

3.2 Standard NMF formulation

A non-negative matrix, is a matrix where all the entries are $x_{ij} \geq 0$ which we also write in matrix form as $X \geq 0$. Let X be a $m \times n$ non-negative matrix. NMF consists on finding a decomposition $X \approx WH$ where W and H are also non-negative matrices with sizes $n \times r$ and $r \times m$ respectively and $0 < r < \min(m, n)$.

Standard NMF can be formulated as an optimization problem as follows:

$$\min_{H \geq 0, W \geq 0} ||X - WH||_F^2 \tag{3.1}$$

This formulation of NMF turn out to be non-convex. However, the formulation reduces to a convex sub-problem when either H or W is fixed. As such, most of the algorithms that solve NMF rely on non-convex optimization methods such as Block

Coordinate Descent [41].

3.3 NMF Properties

Many researchers have analyzed and characterised the properties of the NMF problem. It has been shown that the NMF problem is, in general, a NP-Hard problem [42]. Also, the work [43] showed that non-trivial NMF solutions always exist and also that the NMF problem can be actually formulated as a convex optimization problem. Nevertheless, a practical convex formulation does not yet exist.

Because of NMF being a NP-Hard problem and lack of practical convex formulations, most algorithms solve the more practical non-convex formulation [41]. Those algorithms only ensure convergence to local optima (reaching an stationary point), but with the advantage of doing so in reasonable computational time.

Solving the non-convex NMF formulation is an ill-posed problem and the solution is non-unique [44, 45]. To address this issue, several variants exist that impose additional constraints on the set of possible solutions to select solutions with some additional desired properties [46]. For example, sparsity constraints can be added on either W or H [47].

3.4 Algorithms

Most of the algorithms for solving NMF are based on the property that the standard NMF formulation reduces to a convex sub-problem when either H or W is fixed, following for a two-block coordinate descent scheme [37]. The method is shown in Algorithm 1. In fact, the sub-problem is a non-negative least squares problem (NNLS) when using the Frobenius norm [48].

Algorithm 1: Two-Block Coordinate Descent

Input: A non-negative matrix X and factorization rank r
Output: Non-negative matrices W and H
Initialize the non-negative matrices W and H
while stopping criterion not met **do**
    Update W by solving the sub-problem $\min_{W \geq 0} ||X - WH||_F^2$
    Update H by solving the sub-problem $\min_{H \geq 0} ||X - WH||_F^2$
end

Inside an iteration, there is no special reason to first fix H and update W and then fix W and update H. The updates could be done the other way around. The stopping criterion is usually a maximum number of iterations and some convergence test based on the improvement of the solution between iterations. Also, the Frobenius norm can be changed to any other cost function.

Multiplicative Update Rules

The usual iterative gradient descent algorithm using additive rules would update the matrices W and H with negative values. To avoid this problem, Lee and Seung presented an algorithm based on multiplicative update rules (MUR) and proved their convergence to an stationary point [39].

The multiplicative update algorithm became popular given its simplicity and ease of implementation [49]. Furthermore, those update rules generalize easily to any beta-divergence cost function [50]. The MUR can be viewed as an adaptive rescaled gradient descent algorithm [41]. In fact, this optimization approach is similar to the Expectation–Maximization (EM) algorithms. The MUR follow the two-block coordinate descent scheme.

The updates rules are the following ones:

$$W_{ij} \leftarrow W_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}} \tag{3.2}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^T X)_{ij}}{(W^T WH)_{ij}} \tag{3.3}$$

However the MUR have some drawbacks. First of all, the algorithm is known to be rather slow [41] compared to other more complex ones. This is in part because it has a first-order convergence rate [51, 52]. Moreover, the multiplicative nature means that the algorithm has trouble updating zeros or very small values in the matrices. Nowadays, other more sophisticated algorithms exist to solve NMF, such as Hierarchical Alternating Least Squares (HALS) algorithm [53].

3.5 Choosing the number of components

The number of components (basis or sources) to be extracted is determined by the rank r. The rank r is a key NMF parameter. It fixes the dimensions of the factor matrices H and W, thus, deciding how many sources are extracted. It should be noted that the

standard NMF formulation is not sequential like SVD or PCA, meaning that the factor matrices using rank r can be completely different to those of rank $r + 1$ [54].

There are several common methods used for choosing the rank r. The obvious one is to try different values of r and choose the one performing the best for the application problem at hand. The model order can be robustly estimated with cross-validation scheme [55] when we have a classification task at hand. Moreover, the rank r can be estimated by looking at the decay of the singular values of the input data matrix [56]. Other approaches in the literature involve using a Bayesian approach [57] or meta-heuristics.

One common heuristic for the quality of the obtained NMF solution is to look at how much of the data matrix X variability is retained in the approximation. We can compute a *Relative Reconstruction Error* as $\frac{||X-WH||_F}{||X||_F}$. Naturally, the lower the relative reconstruction error, the better.

3.6 Initialization

The solution obtained by the NMF algorithms will depend strongly on the initialization, given that different initializations will converge to different local optima. In the literature, there are many ways of initializing the initial matrices W and H.

3.6.1 Random

The most simple method is to initialize the matrices with random non-negative values drawn from a known statistical distribution, such as the uniform distribution.

3.6.2 NNDSVD

Perhaps one of the most popular methods is *Nonnegative Double Singular Value Decomposition* (NNDSVD) [58]. As explained on the *Nimfa* library [59], NNDSVD is a method to enhance the initialization stage of the NMF when sparsity is desired. The method has no randomization and is based on two SVD processes: one approximating the data matrix and the other approximating positive sections of the resulting partial SVD factors.

If sparsity is not desired, there is the variant NNDSVDa which fills the zeros with the average of the data matrix X, and the faster variant NNDSVDar which fills the

zeros with values drawn from a uniform distribution.

NNDSVD is well suited to initialize NMF algorithms with sparse factors. Furthermore, empirical evidence suggests that NNDSVD leads to a fast reduction of the approximation error in many NMF algorithms [58].

3.6.3 Random Vcol

Random Vcol [60] initializes each column of the basis matrix W by averaging m random columns of the data matrix X. The initialization of encoding matrix H is performed in a similar way but row-wise. The reasoning behind the procedure is that it can make more sense to build the basis vectors from the data rather than randomly.

3.6.4 Random C

Random C [60] initializes each column of the basis matrix W by averaging m random columns at random from the columns in the data matrix X with largest ℓ_2 norm. Therefore, the most dense columns of the data matrix X are used. The initialization of encoding matrix H is performed in a similar way but row-wise.

3.7 Cost function and beta-divergences

The most popular cost function is the Frobenius $||\cdot||_F$ norm but it can be changed to any beta-divergence cost function $d_\beta(X, WH)$ [50]. The beta-divergence distance can be defined as [61]:

$$d_\beta(X,Y) = \sum_{i,j} \frac{1}{\beta(\beta-1)}(X_{ij}^\beta + (\beta-1)Y_{ij}^\beta - \beta X_{ij}Y_{ij}^{\beta-1})$$

being the Frobenius norm when $\beta = 2$. For $\beta = 0$ and $\beta = 1$ the defined beta-divergence is not continuous. However, these two special cases can be continuously extended and are known as Kullback-Leibler (KL) and Itakura-Saito (IS) divergence respectively [61].

$$d_{KL}(X,Y) = \sum_{i,j}(X_{ij}\log(\frac{X_{ij}}{Y_{ij}}) - X_{ij} + Y_{ij})$$

$$d_{IS}(X,Y) = \sum_{i,j}(\frac{X_{ij}}{Y_{ij}} - \log(\frac{X_{ij}}{Y_{ij}}) - 1)$$

NMF with the Kullback–Leibler (KL) divergence cost function has been shown to be equivalent to probabilistic latent semantic analysis [62, 63].

3.8 NMF Variants

3.8.1 Sparse NMF

Sparse NMF is a NMF variation which enforces sparseness in either the basis matrix W or the encoding matrix H [47]. If sparseness is enforced on the left matrix W the formulation is called SNMF/L and if spareness is enforced on the right matrix H then its called SNMF/R. Those formulations minimize the ℓ_1 norm [47].

Depending on which matrix the sparseness is enforced in, we can encourage the learning of local or global solutions. This allows to enforce the learning of localized sparse parts-based representations, or the opposite, which would yield a global solution, that is, learning prototypes of objects instead of parts.

3.8.2 Semi-NMF

Semi-NMF is a variation of NMF which does not restrict the signs of the data matrix X and the encoding matrix H [64]. This extension is motivated from the perspective of clustering for when the data matrix X has negative values.

3.8.3 Convex NMF

Convex NMF is a variation that, like Semi-NMF, does no restrict the signs of the data matrix X but with the additional constraint that the matrix H has to be formed as a convex linear combination of the data matrix X. So the factor matrix H has do be of the form $H = XG$ for some matrix G [64].

3.8.4 Separable NMF

Under the separability assumption (equivalent to the pure-pixel assumption), the resulting modified NMF formulation can be solved in polynomial time [65] and it is known as Separable NMF. The pure-pixel assumption for images is the idea that for each feature, there is a pixel that is exclusively used by that feature.

3.8.5 Nonnegative Matrix Underapproximation

There is a NMF variant called Nonnegative Matrix Underapproximation (NMU) [66] which extracts sequentially the sources. It has been shown to lead to sparse solutions [67].

3.9 Other factorization methods

3.9.1 Independent Component Analysis

Independent Component Analysis (ICA) is another approach to matrix factorization where the data instances are assumed to be independent and not Gaussian [68]. The algorithm is usually used to separate a signal into additive components that maximize statistical independence. It can also be used as a factorization method to learn a decomposition with some sparsity. Since the formulation of ICA does not include a noise term, whitening of the data is applied [69].

3.9.2 Singular value decomposition

In linear algebra, the Singular Value Decomposition (SVD) is a factorization of a matrix into three matrices.

Given a matrix X of size $n \times m$, SVD can be formulated as an optimization problem as follows. Let $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $\Sigma \in \mathbb{R}^{r \times r}$ a diagonal matrix.

$$\min_{U,V,\Sigma} ||X - U\Sigma V^\top||_F^2$$

subject to $U^\top U = I$, $V^\top V = I$ and $\text{diag}(\Sigma) \geq 0$.

If we center the data, then we are performing PCA on a co-variance (correlation if data is standardized) matrix of the centered data X.

If we have missing values, the problem can solved as a weighted PCA, with the missing values having weight zero. This approach is also called Robust PCA.

3.9.3 Vector Quantization

Vector Quantization (VQ) learns a basis consisting of prototypes, each of which is a representative individual of the data. One typical algorithm to solve VQ is the k-means clustering algorithm.

3.10 Relationship between methods

The NMF objective is to find a product of matrices WH which minimizes the reconstruction error of X. The most popular loss is the Frobenius norm $||\cdot||_F$. This norm is popular thanks to its properties, and equals to assuming the noise on the data to be Gaussian [37].

Furthermore, the assumptions of the noise or structure of the matrices W and H, yields different methods [70]. For example, if we solve the matrix factorization problem constraining the basis W to be orthonormal and the encoding H to be orthogonal, the optimization problem becomes equivalent to SVD [39]. If additionally we assume the data to be centered then the problem is equivalent to the well known PCA.

Alos, when optimizing the Frobenius norm, the NMF, Semi-NMF and Convex NMF can be seen as relaxations of k-means clustering, also called *soft* k-means [71, 72]. Thus, NMF can be motivated and interpreted as a more flexible clustering algorithm.

Chapter 4

Background concepts

In this chapter, we introduce several miscellaneous concepts from statistics, machine learning (ML) and computer vision. Those concepts will help to understand some of the methods and decisions of later chapters of this thesis. Namely, we explain concepts related to training a model robustly, such as data partitioning and cross-validation. In this chapter, we also explain some types of image transformations such as morphological transformations and image filters. We close the chapter with a brief explanation of some classification methods.

4.1 Data partitioning

A common theme in ML consists in partitioning the data in a training set and a test set. Usually, the partitions are stratified, that is, the splits are done such that they preserve (approximately) the same proportion of classes (of the target variable) in the complete set. Simply put, stratifying means that the class proportions are maintained in all partitions. It is strongly recommended to shuffle the data and do stratified sampling to obtain a good partitioning.

This data partitioning scheme is used mainly to overcome a problem called overfitting. If a model with enough flexibility is trained on the complete data, it will end learning too much of the particularities of that data sample, and therefore failing to predict reliably future unseen observations. In this scenario, we would say the model overfitted the data. Figure 4.1 displays a visually representative example.

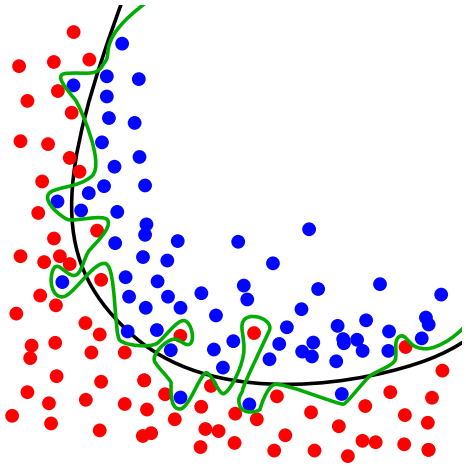


Figure 4.1: The green line is to dependant on the training data, therefore, being likely to have a higher error rate on new unseen data, compared to the black line [73]

By fitting the model with the training data, we can then test it against the test set which the model has yet never seen. With that, we estimate the generalization error more reliably, so that we can get an honest estimate to the real performance of the model.

Sometimes, the data partitioning is done in three sets, training, validation and test set. When done like this the idea is usually to use the train-validation set to fine-tune the hyper-parameters and the test split to estimate the generalization n error.

4.2 Cross-validation

The learning algorithms usually have some parameters that need to be provided by the user and not estimated by the model. Those parameters provide the method with enhanced modeling flexibility, allowing it to work for many different problems. Those parameters are usually called hyper-parameters to differentiate them from the parameters which are those estimated by the model.

However, those hyper-parameters need to be optimized for the task at hand. The usual way to do it is by trying several sensible values. Again, if this procedure was done directly with all the data, we would overfit those parameters to the available data.

To robustly fine-tune the parameters and select the better performing ones in general, a cross-validation (CV) scheme can be performed. The CV procedure consists in shuffling the data and partitioning it into k splits (also called folds) $F = \{f_1, f_2, ..., f_k\}$. For that reason, it is often called k-fold CV. A typical value for the number of folds is $k = 10$.

For each of the k folds, the model is trained over the other $k-1$ folds and its performance is tested against the fold not used for training. There are $\binom{k}{k-1}$ combinations which result in k iterations. The obtained validation errors are usually averaged to get the mean validation error. The idea is to select the value of a hyper-parameter that gives the minimum mean validation error. Having robustly selected the best value for a hyper-parameter, the model is then retrained over all the training data. The image in Figure 4.2 illustrates the procedure.

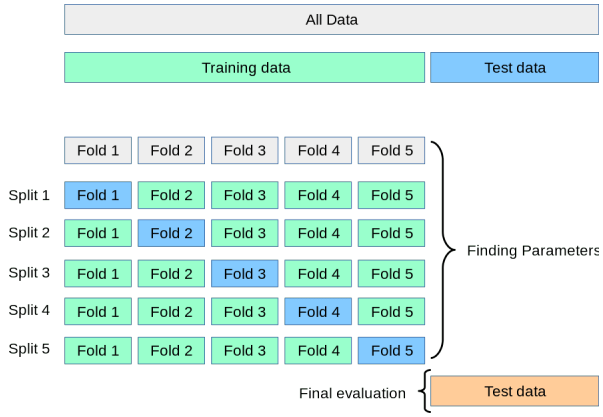


Figure 4.2: Data partition in train-test and 5-fold CV [74]

If the folds are split in a stratified manner then the procedure is usually called stratified k-fold CV. The drawback of CV is that it can be computationally expensive. Nevertheless, it does not waste as many data as, for instance, when fixing an arbitrary validation set.

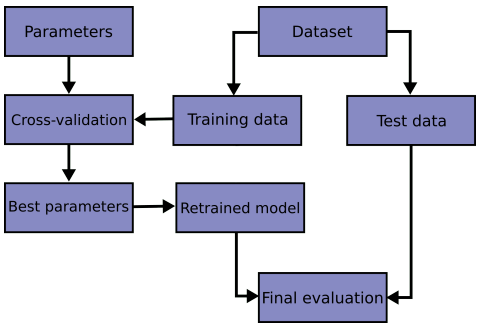


Figure 4.3: Flowchart of typical cross-validation workflow in model training [74]

4.3 Nested CV

Nested CV is an extension of the usual CV scheme. Nested CV allows the estimation of the generalization error of the underlying model while also tuning its hyper-parameters.

In the usual partitioning of the dataset, one fixed split is arbitrary chosen as test data. Nested CV generalizes the concept of the train-test partitioning. Instead of choosing an arbitrary test partition, an outer CV loop splits the data in k_{outer} folds. Like in regular CV, in each iteration one fold is used for testing and the remaining ones are used for training. However, in this case an inner CV is performed on the training partition, further splitting it on k_{inner} train-validation folds. For each train-validation split, a model is fitted on the train fold for a set of hyper-parameters and the performance estimated on the validation fold. The set of hyper-parameters that maximizes the averaged validation score are selected. A model is re-trained for each train-validation split with the best hyper-parameters and then tested on the outer CV test fold to estimate the generalization error [74].

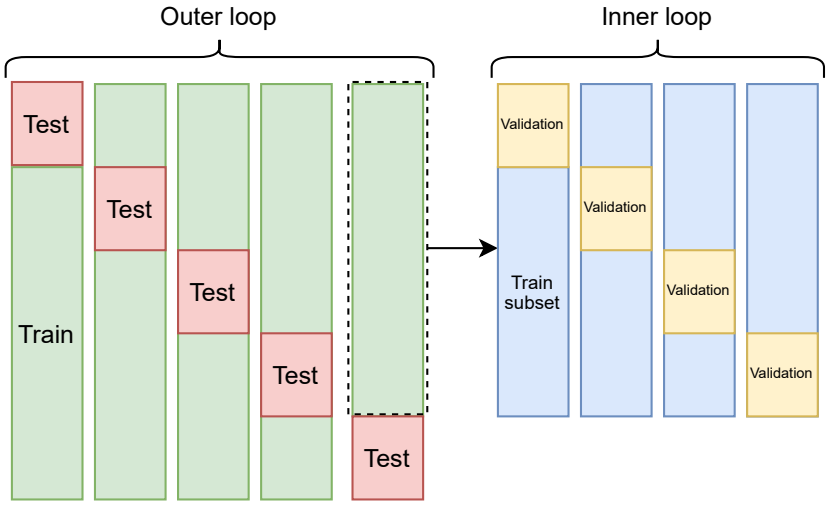


Figure 4.4: Double cross-validation illustration with outer 5CV and inner 4CV

In essence, nested CV uses a series of train-validation-test set splits to tune the hyper-parameters and at the same time get a robust estimation of the generalization error.

In fact, nested CV can be seen as repeating the training of the model each time with a different train-test split. In a way, it is kind of repeating "an experiment". The important invariant is that the test split is always unseen data for the trained model so that it can be used for estimating the generalization error.

The drawback, again, is that it is computationally expensive (more than regular

CV). Nevertheless, this way we can get a more robust *honest* estimation of the performance of the model while also optimizing the hyper-parameters.

4.4 Image transformations

In this section, we review some basics of image transformations, such as image thresholding, morphological transformations, or image filters.

4.4.1 Morphological transformations

Morphological transformations are simple operations applied usually on binary images based on the image shape. The operation takes as input the image and a structuring element (also called kernel). The kernel determines the neighborhood to be examined around each pixel where the operation is applied [75].

The operation consists on sliding the kernel through the image (like a 2D convolution would). Therefore, the shape of the kernel (structuring element) allows control over which shapes we want to affect with the operation. The default kernel is usually a square.

The two most basic operations are called *Dilation* and *Erosion*. Erosion decreases the white boundaries of the image shape. In other words, the thickness of the white region decreases in the image. Dilation in the opposite transformation, it increases the white region in the image. If we apply erosion followed by dilation, then we remove noise (white speckles) in the image. This operation is called *opening*. If we instead apply dilation followed by erosion, the operation is called *closing* and it is useful in closing small (dark) holes in white objects. In Figure 4.5, an example of each operation is displayed.

Figure 4.5: Examples of morphological operations taken from [75]

There are some other transformations, such as *morphological gradient*, *top hat* and *black hat*. The OpenCV library provides a great guide on the topic [75].

4.4.2 Image Thresholding

Image thresholding is the simplest method to segment an image and get a binary mask. As the name implies, a threshold value is set, and all pixel values smaller than the threshold are set to 0, and 1 otherwise. More sophisticated methods of image thresholding exist. There are automatic methods to select the threshold value, being *Otsu's method* one of the most famous. Also localized methods exist, which use a different threshold for each small region in the image. Moreover, the thresholding methods can be generalized to output a grayscale gradient transition instead of a binary mask [75].

4.4.3 Image Filters

There exist many several filters that can be applied to an image. One interesting family of filters are the blurring methods to reduce noise in images. The image filters can be viewed as applying a two-dimensional convolution to the image. By setting the appropriate kernel and convolving it with the image, many types of filters can be applied [75].

Averaging

The simplest blur filter consists on convolving the image with a square normalized kernel. In essence, this operation averages the pixel values under the kernel area.

Gaussian Blur

One popular blurring method is Gaussian blur, which corresponds to applying a Gaussian function to the pixels of the image. The method works well for removing Gaussian noise on the image.

Median Blur

If, instead, we have *salt-and-pepper* noise, then a median filter is highly effective. Median blur works by taking the median of all the pixels under the kernel area and replacing the central element with the median value.

Bilateral Filtering

Another interesting filter is bilateral filtering, which removes the noise while keeping the edges sharp. The method combines the use of two Gaussian filters to achieve the

result.

Gamma correction

Gamma correction can be used to adjust the brightness of an image by using a non-linear transformation between the input values and the mapped output values. On Figure 4.6 we can see the relationship of input and output value.

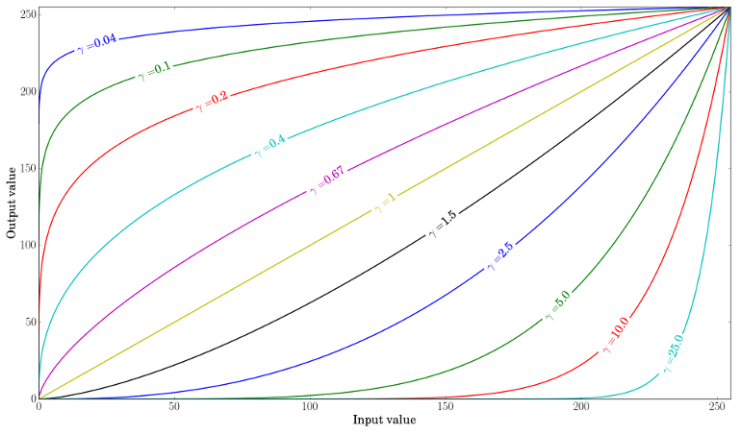


Figure 4.6: Relationship between input and output values for several gammas [76]

Assuming an input image of 8-bit unsigned depth (0 to 255), gamma correction corresponds to applying $Output = \left(\frac{Input}{255}\right)^{\gamma} \times 255$ for some γ value. As this relation is non-linear, the effect will depend on the original pixel value and will not be the same for all the pixels. When $\gamma < 1$, the original dark regions will be brighter whereas it will be the opposite for $\gamma > 1$.

4.5 Classification Models

A ML method is any algorithm that builds a model learning from data which can then be used on new unseen data to make predictions. Many types of models exist, each one with its use cases. For a detailed, thorough and technical explanation to this topic and other ones, we refer the reader to the book [77]. In this section, we glimpse over some standard well known ML classification methods that will be used in the experimentation. A classification problem consists on identifying which category an observation belongs to.

4.5.1 Logistic Regression

Logistic Regression is an ML/statistical supervised method that fits a logistic function for classifying a set of given instances in usually two outcomes. The method models the probability of an instance belonging to a class. Then, a threshold value (usually 0.5) defines the boundary between both classes. When the probability of an instance surpasses the threshold value, the instance will be predicted as belonging to one class, and otherwise belonging to the other class. The method can be extended to work with multiple classes.

4.5.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised method that works by finding a linear combination of the data features that maximizes the separability of the classes. This separation is achieved by maximizing the distance of the means for each category while minimizing the variation of the classes within a class. The new instances will be predicted by computing the probability of being part of each of the defined classes.

4.5.3 Support Vector Machine

A Support Vector Machine (SVM) is a supervised ML model that finds a decision boundary that splits the data with the highest margin possible. In the basic SVM description, a linear decision boundary is learnt by defining a high dimensional hyperplane to split the data, and usually works well for linear binary classification problems. Additionally, SVMs can also be used for non-linear binary classification problems. The SVMs can learn non-linear decision boundaries using what is called the "kernel trick", implicitly mapping the input data into high-dimensional feature spaces. The mapping is done through a "kernel function", being the most famous the radial basis function (RBF). This allows the SVM to work for non-linear binary classification problems.

Chapter 5

Experiments

In this chapter, we report the experimental settings of the thesis. We explain the methodology, the task of interest, and describe the dataset. Furthermore, we report the model exploration and how we arrived to the final preprocessing of the images. Finally, we present and discuss the results.

5.1 Methodology

A thorough explanation of the methodology is provided in this subsection. Also, the methodology is summarized on the detailed diagram in Figure 5.1.

For each of the image modalities, we will execute the factorization methods to learn an unsupervised representation of the data. There are three image modalities in the dataset, namely FR, OCT and OCTA. For the OCTA images there are four sub-modalities: $3 \times 3mm$ *superficial*, $3 \times 3mm$ *deep*, $6 \times 6mm$ *superficial* and $6 \times 6mm$ *deep*.

The needed filtering, and basic data preprocessing will be carried out. Then, we will start working with small standard NMF models to see how the learnt sources look like and see what features are captured. If the features captured are not interesting, we will try to preprocess the images to remove such non-informative features.

We seek a parts-based representation. With that, we mean a representation where we combine in a additive way some features to build the original image. What this essentially means is that the target image will be reconstructed through a linear combination of some learnt relevant components. In this setting, we will prefer the components to capture sparse localized features so that the parts-based representation is easier to understand.

After the images are filtered and preprocessed, unsupervised LDR models will be built for each of the six types of image. The models included are NMF, Sparse NMF, Separable NMF, Convex NMF, SVD, PCA, Sparse PCA, Factor Analysis, ICA.

We run the models with two approaches: considering each image as an *individual* and considering each pixel as an *individual*. For some of the models, the results and interpretations can be different depending on the approach.

We will estimate the rank r (number of components) by looking at the decay of the SVD eigenvalues. We will take a relatively large number of components so as to not restrict too much the features the models can learn.

For each combination of image type, model and parametrizations, the decomposition obtained is made of components (basis) and weights (encoding). Following the same scheme as other works [21] that use NMF for classification, we will use the encoding matrices (H of size $r \times n$) as input variables for the classification models.

This approach has the benefit of making the combination of different models built for each image type straightforward. If there is at least some patterns captured by sources which discriminate to some degree our task of interest then their weights will reflect that. Still, since there will be a huge amount of weight variables we will perform feature selection prior to robustly training the classifiers.

The measure of interest used to evaluate the suitability of source extraction for the retinal images is the ability to discriminate better than random in the following two binary classification problems.

- Discriminate between the eyes of non-diabetics (controls) and diabetics.
- Discriminate between asymptomatic (no DR) and symptomatic (DR) eyes.

To compare the performance of the models, the *Area Under Curve* (AUC) metric will be used. This is a common metric used for this type of problems. The AUC metric is defined as the area under the *Receiver Operating Characteristic* (ROC) curve. The ROC curve quantifies the performance of a binary classifier for all the classification thresholds. A classifier which always predict the majority class would have only 1 threshold and an AUC of 0.5. Also, a classifier which predicts following the prior distribution of the training data will average an AUC of 0.5.

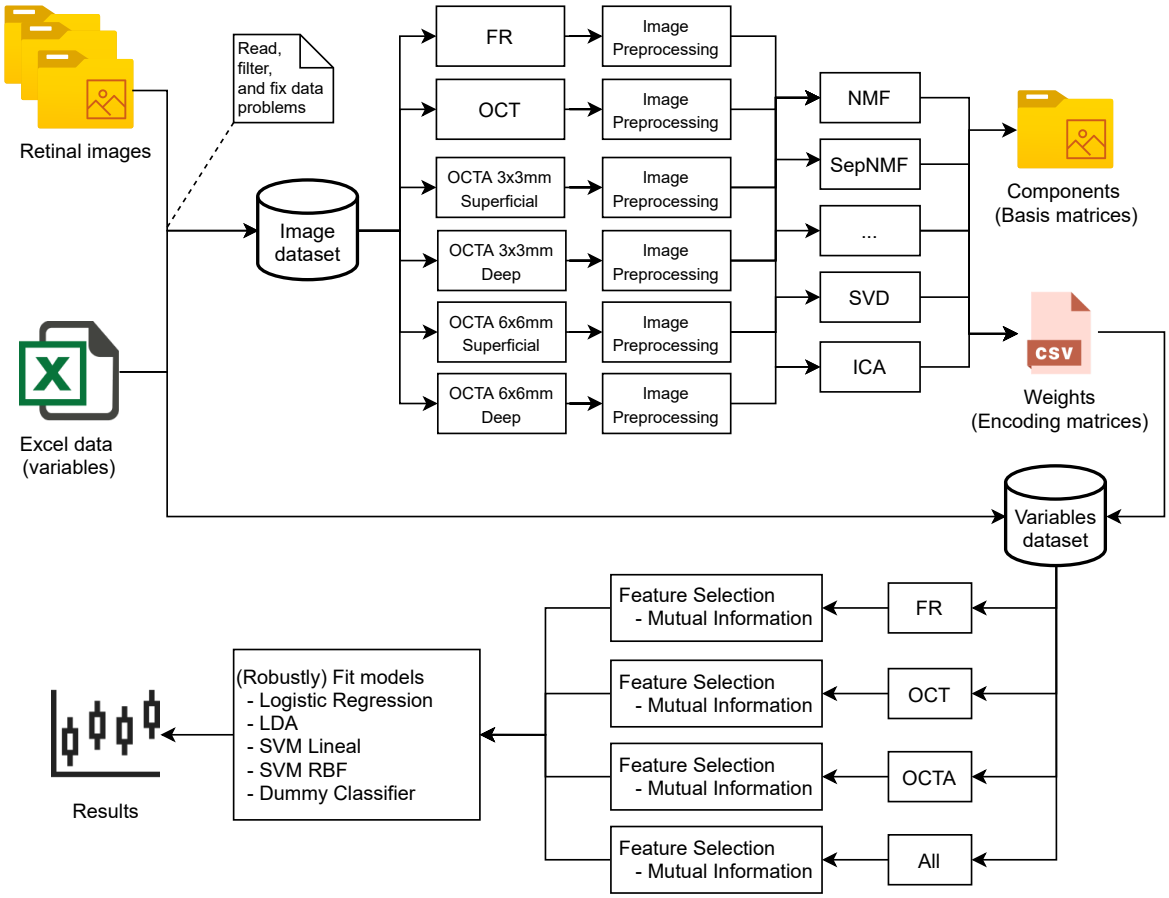


Figure 5.1: Diagram summarizing the project flow

5.2 Tools

For the development of this thesis we use Python 3. The most relevant libraries are the following ones:

- For manipulating the data we use the libraries NumPy and Pandas.
- For manipulating the images we use OpenCV, scikit-image and Pillow libraries.
- For executing the models we use sklearn, nimfa and PyNMF.
- For plotting and diagrams we use matplotlib and seaborn.

Needless to say, other tools were used, such as *LaTeX* to write this thesis or *diagrams.net* to create some of the diagrams.

5.3 Dataset description

The dataset differs from many other datasets of common use mainly in that it has only patients with Type 1 diabetes. Furthermore, in the provided labeled dataset there is an extra class which are the controls, healthy people which do not have diabetes who volunteered to undergo the taking of the images and clinical data. Therefore, for our dataset, the DR scale is redefined to include the controls. On Table 5.1 we can see the redefined scale.

0	Controls
1	No DR
2	Mild
3	Moderate
4	Severe
5	Proliferative DR

Table 5.1: DR scale rating when including controls

It should be noted that it would be difficult for a trained optometrist or ophthalmologist to discriminate between the retinal images of controls (class 0) and diabetics with no DR (class 1). That is because diabetics with no DR (class 1) with a healthy eye would not have any obvious signs on the FR and OCT images. Some recent developments show promising results for OCTA images [78].

The tasks of interest are to see if it is possible to discriminate better than random the following two binary classification problems.

- Discriminate between the eyes of non-diabetics (controls) and of Type 1 diabetics which correspond to class [0] versus class [1,2,3,4,5].
- Discriminate between asymptomatic (no DR) and symptomatic (DR) eyes which corresponds to class [1] versus class [2,3,4,5].

The dataset includes 599 people in total. They have sequential numbered identifiers from 1 to 599 both included. For each of the included people their medical record and clinical history data is available (with missing values). Furthermore, the retinal images acquired with the three imaging techniques (FR, OCT, OCTA) are available for both the left and right eye (whenever possible). Note that this combination of modalities, together with clinical data of the patients is an exceedingly rare combination to find in existing research.

Although the clinical data has interesting variables, they will not be used for classification, as this has already been explored by previous work [35] and because we aim to focus on the use of source extraction.

Of course, for a variety of reasons some of the images (and clinical information) are missing. Therefore, the data will need to be filtered and preprocessed. The first step is to sort out all the nuances and errors in the data. Moreover, we need to make sure to align the data in the excel sheet with the clinical information with the dataset images. Sequential reading of the information will fail, given that there are some people with missing images.

The process of aligning the folder IDs with their corresponding IDs in the Excel sheet was carefully supervised by an expert ophtalmologist. This confirmation was needed because, among other reasons, there is only clinical data for 596 people but 599 folders of information were available. Also, some of the persons have a missing value on the target label variable in the Excel sheet.

Regarding the filter, we also have that some of the OCT or OCTA scans are corrupted. Some examples can be seen in Figure 5.2.

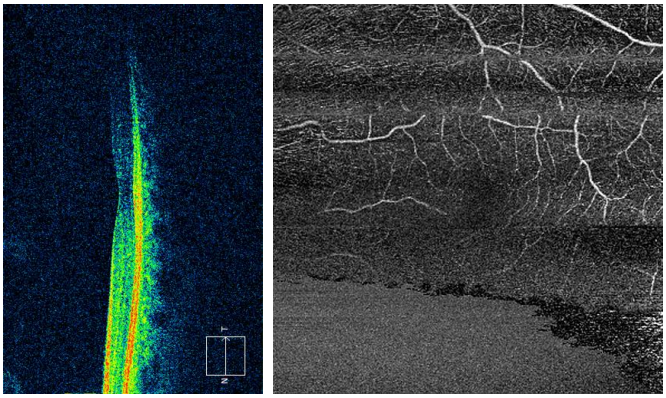


Figure 5.2: Example of corrupted OCT and OCTA

Besides, some of the eyes on the dataset have other eye pathologies or previous treatments. To avoid biasing the model performances, only people with good enough quality OCT and OCTA were included. Also, those users that have been under treatment or surgery that can affect the captured features were filtered out. Following the advice of the ophthalmologist, the filter is implemented as specified on a previous study [35]. The filter is based on variables found on the Excel data and it filters all OCT and OCTA images. The FR images do not have any quality information and so all are included. To be able to make use of all the image types in the dataset without having

missing values, an eye is only included if and only if it is an included eye for all the image types (FR, OCT, OCTA).

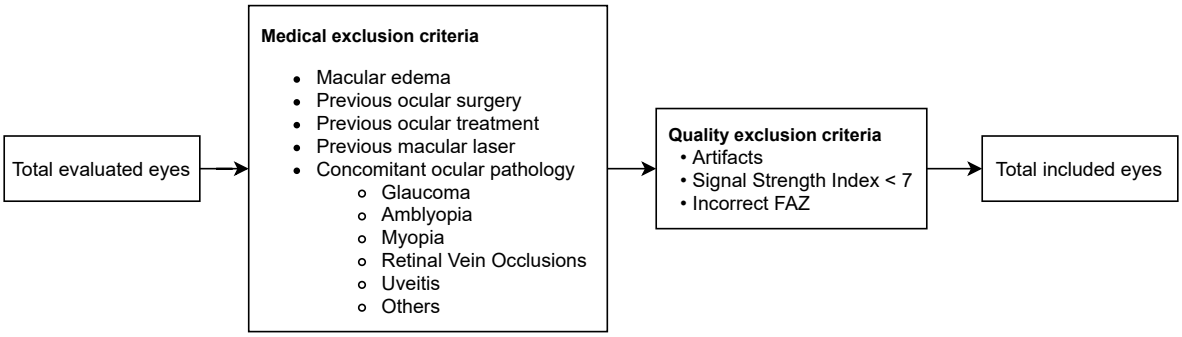


Figure 5.3: Diagram showing the exclusion criteria [35]

After applying the exclusion criteria, we are left with 771 eyes. There was a discussion to consider if we should work with each eye as a individual or join the information for each patient. In the end, the analysis unit was chosen to be each individual eye to have a larger dataset. According to Dr. Zarranz-Ventura, there should not be a bilaterality bias as in a clinical setting joining the information is only required when assessments are done for systemic factors (i.e. age, duration of disease, blood parameters, etc.).

DR scale	No. eyes before filter	No. eyes after filter
0	228	136
1	610	445
2	245	162
3	42	25
4	5	2
5	44	1

Figure 5.4: DR class counts

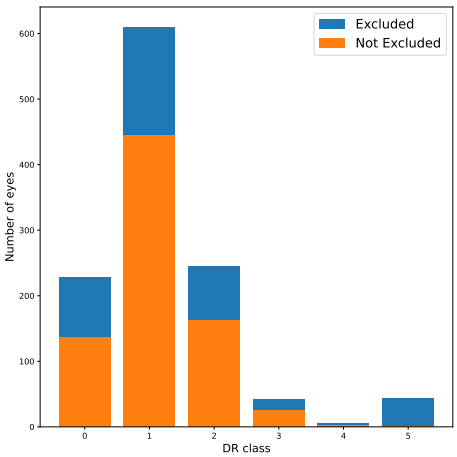


Figure 5.5: Distribution of labels

Looking at the distribution of the instances in the dataset, we noticed that there are too few class 3, class 4 and class 5 eyes after filtering. Thus, it makes more sense to aggregate them into three classes: class 0, class 1 and class [2,3,4,5].

5.3.1 Quality of retinal images

The FR images have dimensions 2576 × 1934 (Width × Height) and are saved in TIF file format. The images are in RGB color space, and the metadata reports 24 bit depth (8 bit per channel) which is saved uncompressed. The file size of each image is 14.2 MB.

The OCT images have resolution 506 × 338 (Width × Height) and are saved in JPEG file format. The images are in RGB color space, and the metadata reports 24 bit depth (8 bit per channel). The file size of each image is around 65 KB. There is one image (0062 Left Eye) which has size 508 × 338.

The OCTA images have resolution 1024 × 1024 (Width × Height) and are saved in bitmap (BMP) file format. The images are gray-scale with 8 bit depth. The file size of each image is 1.0 MB. There are 6 images (from people with identifiers 0179, 0364 and 0577) which have 32 bit depth (8 bit per channel) and weight 4.0 MB accordingly. Inspecting those images and comparing their individual color channels revealed that each image only has one channel but repeated.

There is a folder for each person named with their ID on the dataset. Inside each folder, there are their corresponding images. The image filenames have clear patterns that identify which eye and type of image is, being this the only way for reading them automatically. However, there are some typos and errors. Therefore an analysis was conducted to identify the problematic images. The analysis was done by counting how many images each person has for each image type. Those who did not have the expected counts were manually inspected.

- The folders with ID 0309 and 0593 are empty.
- The people with ID 0035, 0233, 0377, 0396 have a missing right eye.
- The folder with ID 0015 has the left eye missing.
- The folder 0062 has no OCTA $6 \times 6mm$ for the right eye.
- The folder 0102 has no right eye OCTA $3 \times 3mm$ *deep* and no right eye OCTA $6 \times 6mm$ *superficial*.
- The folders 0102, 0103, 0106, and 0137 have wrong formatted filenames which were fixed.
- The folder 0456 has two extra duplicated left eye OCTA $3 \times 3mm$ images.
- Folder 0336 has its right eye OCT wrongly named and an exact copy of the left eye OCT. The wrongly named right eye OCT was removed.

- Folder 0140 has an extra left eye retinography image which was removed.

After all of the applicable changes were performed, we proceeded to convert the dataset to compressed PNG format to save disk space without losing quality (original dataset has size 21.4 GB). Also, to avoid the need to always apply pattern matching on the names, it made sense to separate the dataset for each of the images types.

5.4 Exploration

The first exploration step consisted on experimenting with small models for each type of image and then inspect the extracted components. This way, we can have a first insight on what models learn and how they can be improved.

5.4.1 Retinography

With the FR photographs we built some small models extracting a few sources and look at what the model is learning. We started by learning from the images in a RGB (Red, Green, Blue) color space.

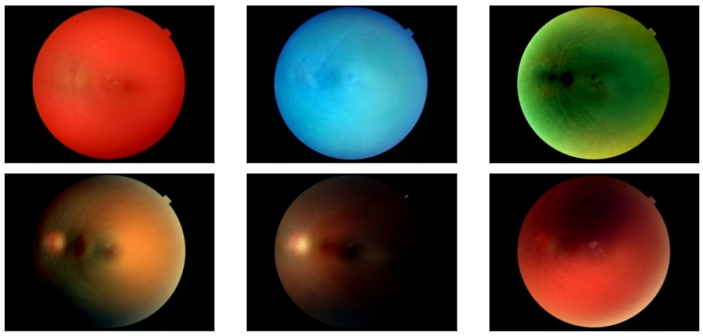


Figure 5.6: Six sources of the FR exploratory model

We can see that models are learning to separate the color channels, the shades and illuminations, which is not what we are looking for. The goal is to normalise the illumination, which can be achieved with local adaptive filters. Regarding the color channel, we know that the vessels are mostly on the green channel [79]. Furthermore, the winner of the Kaggle 2015 competition made public his preprocessing of the images, so we will try to use a similar approach.

5.4.2 OCT

Starting with the OCT scans we build small models using few sources and look at what the model is learning. We started learning the images in a RGB (Red, Green, Blue) color space.

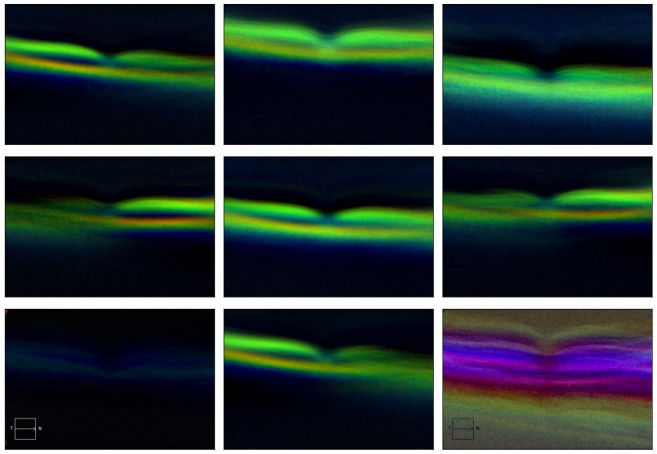


Figure 5.7: Nine OCT sources of the exploratory model

Looking at the results on Figure 5.7, we can see that the model is actually learning different translations and rotations of the OCT scans, which is not what we sought. Moreover, the last source, which looks quite strange, it is actually due to some OCT scans being in a grayscale color space instead of RGB.

In fact, it was confirmed with the ophthalmologist that the OCT scans were originally grayscale and that the color in the image is just an added color map for visualization. Therefore, we converted all images to grayscale.

Furthermore, to be able to learn more intrinsic features, instead of the obvious ones such as position and rotation, we need to develop a preprocessing step to isolate the region of interest (ROI).

Another thing we noticed is that there is a legend and symbol on the left of the OCT scan. This is recurrent in (almost) all of the OCT scans and it is isolated in its own source by the model, as can be seen in the bottom left source of Figure 5.7.

In fact, for the OCT images in grayscale, NMF was executed with $r = 1, 2, ..., 50$ to inspect the behaviour of increasing the rank r. The pattern superimposed over the OCT-scans (the magenta bar and the legend of the OCT-scan) which is placed almost in the same place for all images, was not factorized until rank $r = 13$.

Figure 5.8: Legend and lateral bar extracted by the NMF decomposition

5.4.3 OCTA

OCTA images are mostly fine as they stand and can be used as is. In Figure 5.9, some of the learnt sources are displayed. Nevertheless, we will consider some noise reduction filters such as a median filter, bilateral filter and different types of image thresholding.

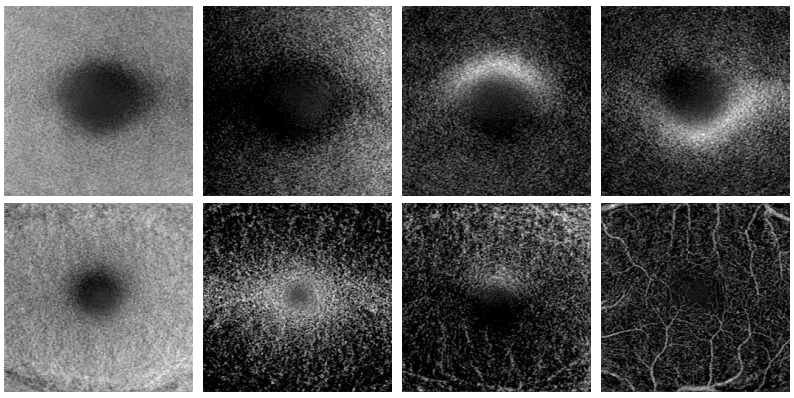


Figure 5.9: OCTA *deep* sources on the above row and OCTA *superficial* sources on the bottom row

Also, when learning, the model identified artifacts present in some of the images. For example, one component helped identify six images which had the camera model watermark on the bottom right of the image.

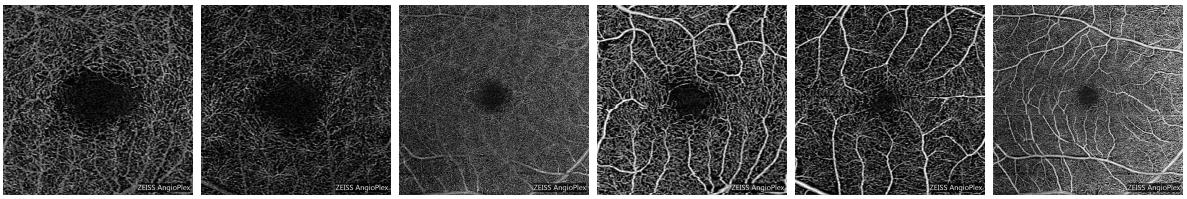


Figure 5.10: OCTA images with a watermark on the bottom right

It also identified the eye of a patient with a very unusual path of the eye nerve through the FAZ area. We can see it in Figure 5.11.

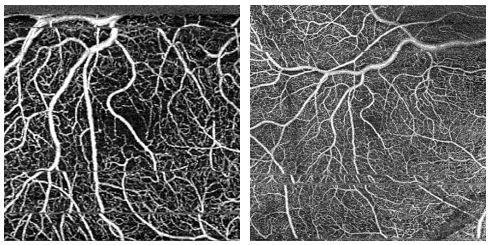


Figure 5.11: Eye with unusual vessel pathing

5.5 Preprocessing

5.5.1 Retinography

We consider here the preprocessing method proposed by Ben Graham, the winner of 2015 DR Kaggle Competition [80]. He preprocessed the images according to the following steps:

1. rescale the images to have the same radius (300 pixels or 500 pixels)
2. subtracted the local average color; the local average gets mapped to 50% gray
3. clipped the images to 90% size to remove the boundary effects

Those preprocessing steps are intended to remove some of the variation on the images due to lighting conditions, exposure, etc. Graham provides an OpenCV implementation in Python 2.7 [80]. In Figure 5.12, we see an example of Graham's preprocessing.

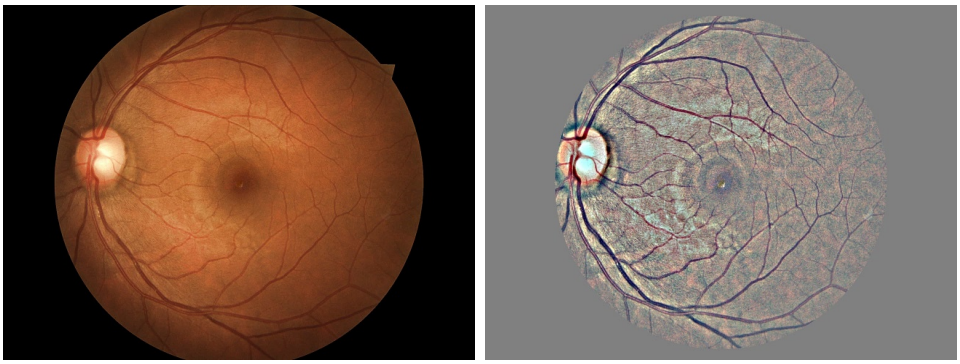


Figure 5.12: Example of Graham's data preprocessing

It should be noted that Graham's preprocessing procedure was designed to be used for a sparse neural network [81]. Using the preprocess "as is" for NMF is not really adequate. To make it easier for the NMF model, the background should remain dark and the vessels should be bright. That encourages NMF to learn vessel patterns or systematic retinal defects. Also this way, the same component can be more easily re-used to reconstruct several eyes.

To achieve such result, the subtraction of the local average color is mapped to white and inverted after that. Moreover, the result is a RGB image for which the red channel mostly captures the hemorrhages and lighting artifacts, the green channel captures the vessels and the blue captures mostly nothing. We can see that in Figure 5.13.

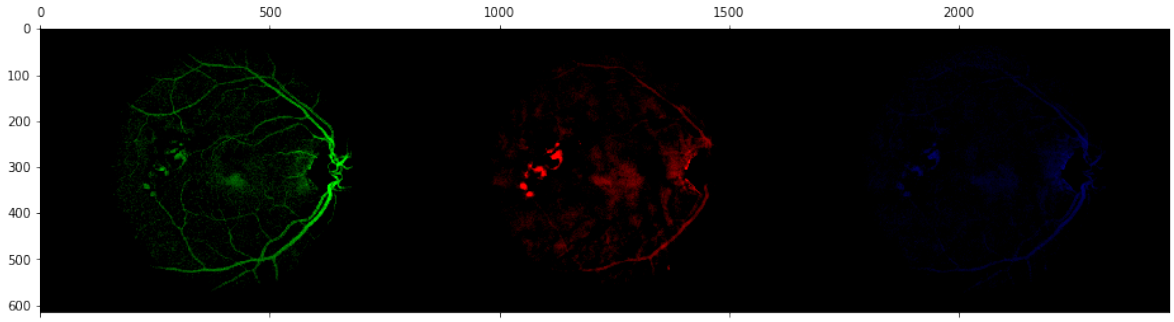


Figure 5.13: Modified preprocessing result shown separately for each RGB channel

Gamma correction was applied on the red channel to leave only the brightest colours. On the green channel, we do the opposite in order to emphasize the vessels.

We also tried to see the suitability of the polar transformation, given that the eye has circular shape. Moreover, the *Fast Fourier Transform* was also explored. Preliminary basic testing did not show promising results. Also, to make the learnt features more robust, we tried splitting the image in small patches and learn from those. Although promising, we did not pursue it because it would be better to have a model capable of learning such localized transformation by itself.

5.5.2 OCT

For OCT, we need to normalize the position and rotation of the images. To do that we crop the region of interest (ROI) and remove the legend and bar in the image.

Overlay

Further inspection shows that all color images have the legend overlaid on the same positions although the few images in grayscale do not. In some cases the overlay is on the region of interest (ROI).

The overlay consists in a magenta vertical line and a white square indicating the orientation of the scan. Since those are very different colors compared to the ones present in the ROI we perform an analysis of a 3D scatterplot where we can look at which range of values are those specific colors located an remove them. An example can be seen on Figure 5.14.

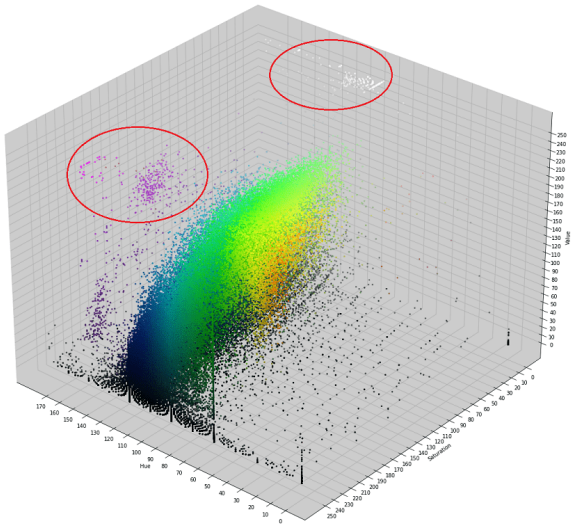


Figure 5.14: 3D scatterplot of an OCT image in HSV color space with highlighted white and magenta colors

We could do that in the RGB space but it was a bit better to perform the operation in HSV (hue, saturation, value) color space. We can see the result of the separation in Figure 5.15.

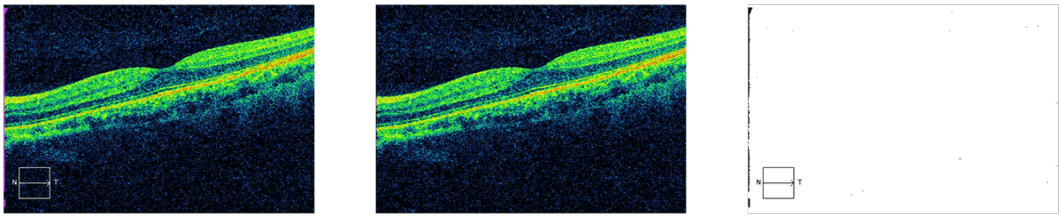


Figure 5.15: Original image, filtered image and mask respectively

Segmentation

After removing the overlay, we explored segmentation methods. We started with Chan-Vese automatic segmentation as seen in Figure 5.16. Although it works very well in most cases, sometimes it crops half of the image. Furthermore, it is a rather slow method.

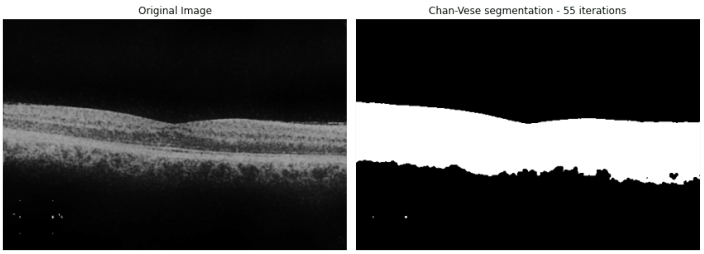


Figure 5.16: Example of Chan-Vese automatic segmentation

Subsequently we proceeded to develop a custom automatic pipeline to extract the region of interest (ROI). We experimented many different methods and ways, by trial and error. In the following section, we explain the final result, skipping most of the failed transformations.

To fix the rotation of the image, we extract an approximation of the shape of the image. To efficiently obtain the shape in a reliable way, we apply a median blur filter followed by a bilateral filter to reduce the noise. The reason is that OCT scans have a significant peak signal-to-noise ratio. To the result, we apply a morphological opening with an horizontal line kernel which has the effect of reducing the vertical lines (while preserving the horizontal). Then the resulting shape is dilated with a vertical line kernel. This procedure extracts the shape of the OCT scan as a binary mask.

Rotation

Once we have the shape of the ROI, we explore how to fix the rotation. PCA was used to find the orientation of the shape. Having the two first principal components of the mask, we can compute the angle between the x-axis and the first component eigenvector with some basic geometry. Given an eigenvector (x, y), we compute the angle as $\arctan(\frac{y}{x})$. The angle is adjusted such that the minimum needed rotation is performed (clockwise or counterclockwise).

Although PCA works well, we found a more efficient way. We can estimate a best fitting ellipse (using the second order moments) and use its properties to get the orientation. This turns out to be more efficient and was done with the *regionprops* function of the scikit-image library.

Figure 5.17: From left to right: *minAreaRect*, *PCA*, *regionprops*

In the end, though, for both rotating and cropping the area of interest, it ended being better to find the best fitting rotated rectangle using the OpenCV library function *minAreaRect*. After getting the coordinates of the rotated rectangle, we mapped it using a perspective transform to a non-rotated rectangle.

Curvature

For people with a bit of myopia, the acquired OCTs are curved. According to Dr. Zarranz-Ventura, the curvature is related to the "*Axial Length*" variable. People with too large curvature are excluded, but that still leaves eyes with some non-negligible curvature.

To straighten the image, we made again use of the mask shape of the image. We know that the curvature will be either convex or concave. Therefore, we fitted a parabola (polynomial of degree 2 or also known as quadratic polynomial) to the mask in order to estimate its curvature. Then, we shifted (also called roll) the columns so that they align with the fitted parabola. An example can be seen in Figure 5.18.

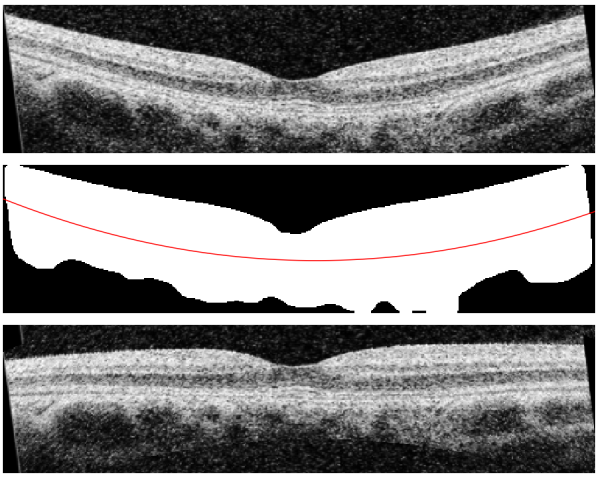


Figure 5.18: Procedure of straightening the OCT scan

Skewness

This perspective transform of the *minAreaRect* rectangle adds skewness to the image. The larger the required rotation, the more skewed is the resulting image. We can see an example on Figure 5.19.

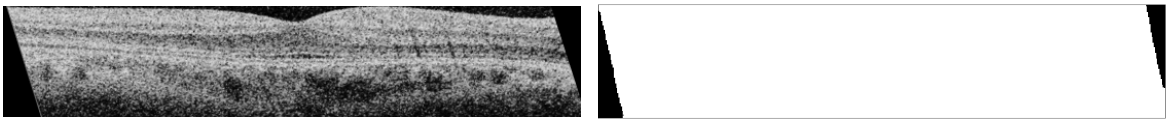


Figure 5.19: Left: the skewed OCT. Right: skewed OCT mask.

To fix the skewness, we obtained a binary mask and used (for efficiency) the second order moments to compute the skewness of the image in the vertical axis. The binary mask can be seen in Figure 5.19.

The skew is computed using the second order moments of the image as $\frac{\mu_{11}}{\mu_{02}}$ and a perspective transform (shear matrix) is applied. In Figure 5.20 we can see the result after deskewing and the final result after cropping the black borders.

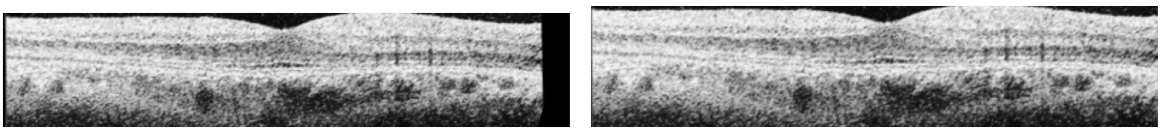


Figure 5.20: Left: the deskewed OCT. Right: final cropped OCT.

As a final step, we applied a non-aggressive bilateral filter to reduce the noise. Since after cropping and transforming, we end up with images with slightly different dimensions, we looked at the mean and median sizes of all the cropped images and used it to make an informed decision to resize the image to size. The results showed that resizing to dimensions 500×100 worked well.

5.5.3 OCTA

The OCTA images are good enough as they are. Nevertheless, for the *superficial* images, we tried some vesselness known image filters such as *frangi*, *meijering* and *sato* [82]. Those filters can be used to detect continuous ridges such as vessels. An example is provided in Figure 5.21.

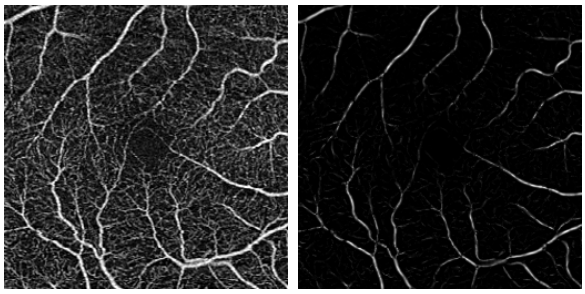


Figure 5.21: Example of applying *frangi* filter to an OCTA $3 \times 3mm$ *superficial*

5.6 Learning unsupervised representation

After having performed all the filtering and preprocessing of the data, we executed a final definitive run for all the models. Each NMF and NMF variant model was run with different initializations. Namely, we ran them with random, NNDSVD and NNDSVDa initializations. The other factorization models (SVD, PCA, etc) were run with the default parameters.

Regarding the number of components to extract in the factorization models, since the task at hand is a classification problem, we could use a CV scheme. However, we would have to tune it for each model and initialization. This is a rather cumbersome procedure and, therefore, we used a different strategy. We decide a sensible range of values to try by looking at the decay to the SVD eigenvalues. This is essentially the same as looking at the retained variance of PCA. We will choose a large enough range and execute for several values of components. We will leave it up to the feature selection to decide which decomposition is the best.

We tried to use the images at full resolution, but it ended being better to reduce the sizes. Therefore, to make to make it more feasible for the factorization models, the size of FR images was reduced to 256×256, OCT to 100×500 and OCTA to 256×256. With this data, we analyze the SVD explained variance, which can be seen on Figure 5.22.

Something interesting to notice is that OCT images have a better curve than the rest. One possible reason could be that the factorization method does not agree well with the vessel variability present on the FR and OCTA images.

Based on the plot, we decided to run the models for $r \in \{64, 128, 256, 384\}$. Going beyond that for most models would result in learning specific individuals, instead of features. For the sparse models though, it could make sense to have even larger number of features, but we do not expect to have huge amount of relevant sparse and localized features in the data.

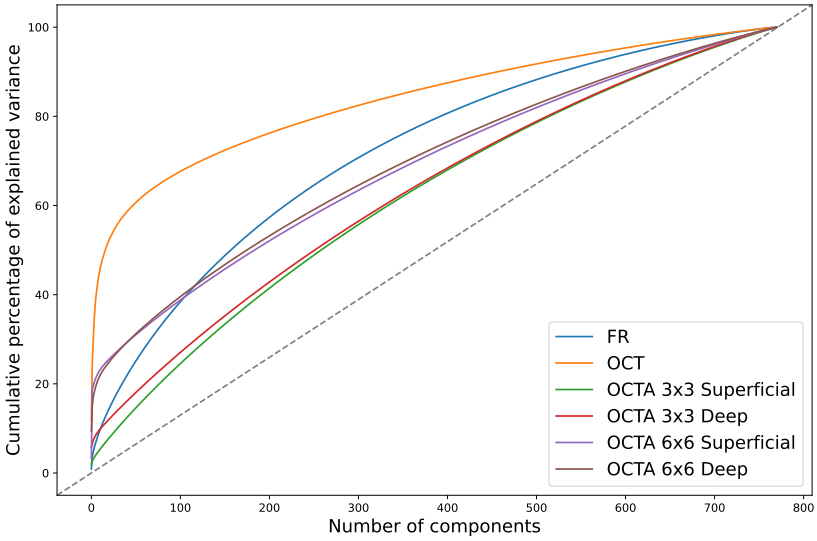


Figure 5.22: SVD cumulative percentage of variance explained

After having learnt the unsupervised representation, the encoding matrices will be used as features. Then a feature selection and double cross-validation scheme will be performed to train and test the models.

5.7 Feature selection and classifier training

The feature selection will be performed using a similar approach as carried out in recent unpublished work with radiomics by student Laura Carrera Escalé under supervison of Dr. Enrique Romero Merino and Dr. Alfredo Vellido Alcacena. A feature selection approach based on mutual information (MI) will be applied and a stratified double-cross CV scheme will be used to robustly train and test the classification models.

The MI between two random variables is a measure of the dependency between the variables. It equals zero if and only if the two random variables are independent, with higher values representing a higher dependency [82, 83].

The feature selection is carried for each of the classification tasks (diabetic or non-diabetic and presence of DR) and for each subset of features (FR, OCT, OCTA and all of them). The following ML/statistical classification methods were used: LR, LDA, Linear SVM and RBF SVM.

Each selected subset of features is ordered from highest to lowest MI with the target class and the first 32 features are selected (the 32 with the highest MI). To do this in a robust way, this procedure is carried out using a 10-fold CV. The MI of each variable is computed for each split. This provides us with 10 MI estimates for each variable,

which are then averaged.

Then, using a double CV scheme, the hyper-parameters of the models are optimized and the generalization error estimated. Specifically, the inner CV is used to select the best hyper-parameters according to the averaged validation AUC metric. Once the parameters of the models have been defined, a (optional) backward elimination wrapper method is applied to remove the irrelevant or less useful features for the model. To check if a feature can be safely removed, the hyper-parameters are re-optimized on the same corresponding inner CV to see if there is a decrease on the averaged AUC metric. Once the features and hyper-parameters are selected, a model is re-trained for each inner CV train split and are tested on the corresponding outer test CV fold.

The stratified double CV is defined with 5 splits on the outer CV and 4 splits of the inner CV. This gives us a total of 20 iterations. Using the explained double CV scheme, we get 20 test estimates that will be shown as a boxplot. We would get 5 test estimates (instead of 20) if we re-train on the outer CV train split.

The grid search for the hyper-parameters of the classification methods is shown in Table 5.2.

Method	Hyper-parameters
Logistic Regression	$C = 10^{-3:3}$
LDA	None
Linear SVM	$C = 10^{1:4}$
RBF SVM	$C = 10^{1:4}$, $\gamma = 10^{-4:1}$

Table 5.2: Grid search values for the hyper-parameters of the classification methods. The notation $x : y$ denotes all the integers in the range $[x, y]$.

5.8 Results

5.8.1 Useful sources learnt

After executing the models, we take a look at the learnt components. We show some of the relevant learnt NMF sources for each type of image when initialized with NNDSVD. We do not show the non-NMF components because those are difficult to interpret.

In Figure 5.23, we can see some of the sources learnt by NMF for OCT images. We can see the sources are a localized parts-based representation.

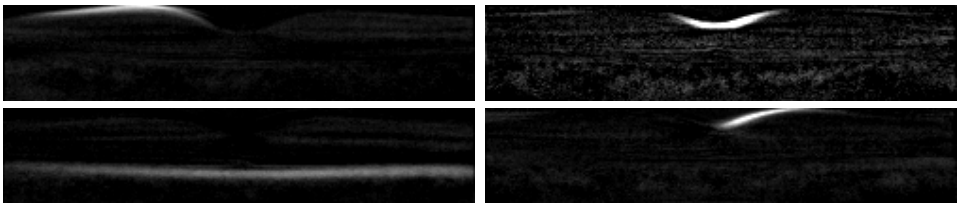


Figure 5.23: Some NMF sources from OCT retinal images (initialized with NNDSVD)

Looking at the learnt NMF components for the *deep* OCTA images (Figure 5.24), we can see that they capture the different patterns around the FAZ area.

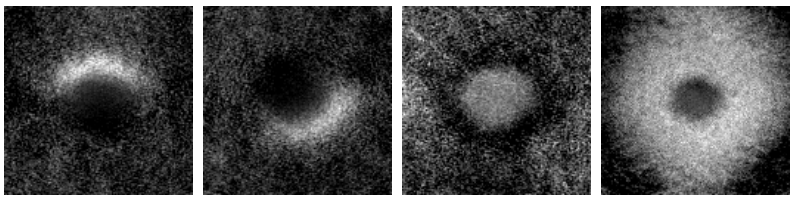


Figure 5.24: Some NMF sources from OCTA *deep* images (initialized with NNDSVD)

The learnt NMF features for the *superficial* OCTA images can bee seen in Figure 5.25. A sparse representation is learnt. We notice that the bottom vessel is being captured by different components depending of its position.

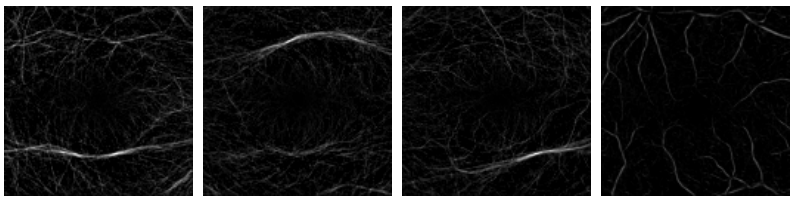


Figure 5.25: Some NMF sources from OCTA *superficial* (initialized with NNDSVD)

Some of the learnt NMF components for the FR images can be seen in Figure 5.26. The sources mostly seem to capture the thickest vessels. Like in the OCTA sources, there is variability on the positioning of the vessels which ends captured in different sources.

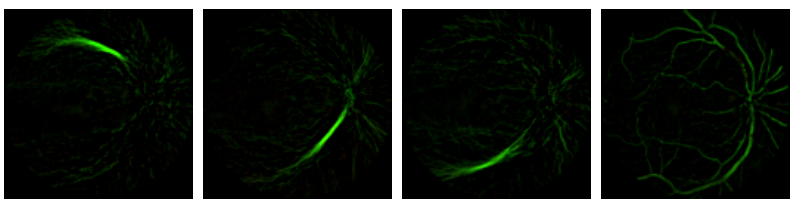


Figure 5.26: Some NMF sources from FR images when initialized with NNDSVD

5.8.2 Classification results

We include a *Dummy Classifier* that generates predictions by respecting the training set class distribution. This classifier will have an average of AUC of 0.5. We note that a classifier which always predicts the class that maximizes the class prior (the most frequent label in the training set) will have an AUC of exactly 0.5.

We tried two approaches, using all the learnt features and using only the NMF and NMF variants features. Even when using all the learnt features, the feature selection procedure ended selecting many of the NMF features as the most useful ones. Thus, the results were pretty much the same in both settings (within a small margin of error). Because of that, the results we present in the following section are those using only the NMF and NMF variants features. This way, we can make better use of the easy interpretability of the parts-based learnt decomposition of the NMF sources.

Discriminating DR

The binary classification task of discriminating between the asymptomatic (no DR) and symptomatic people (DR) corresponds to class 1 versus class [2,3,4,5]. We name this classification task 1-2. For this task, we obtain the results shown in Figure 5.27.

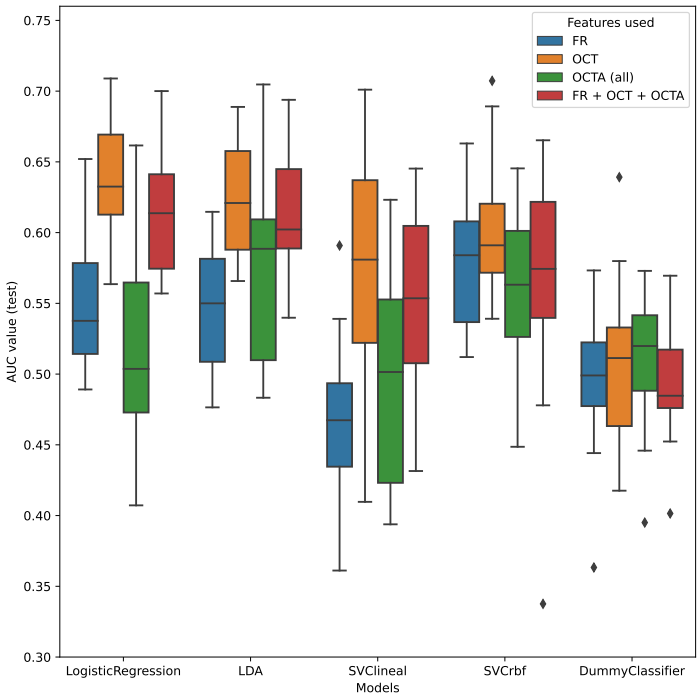
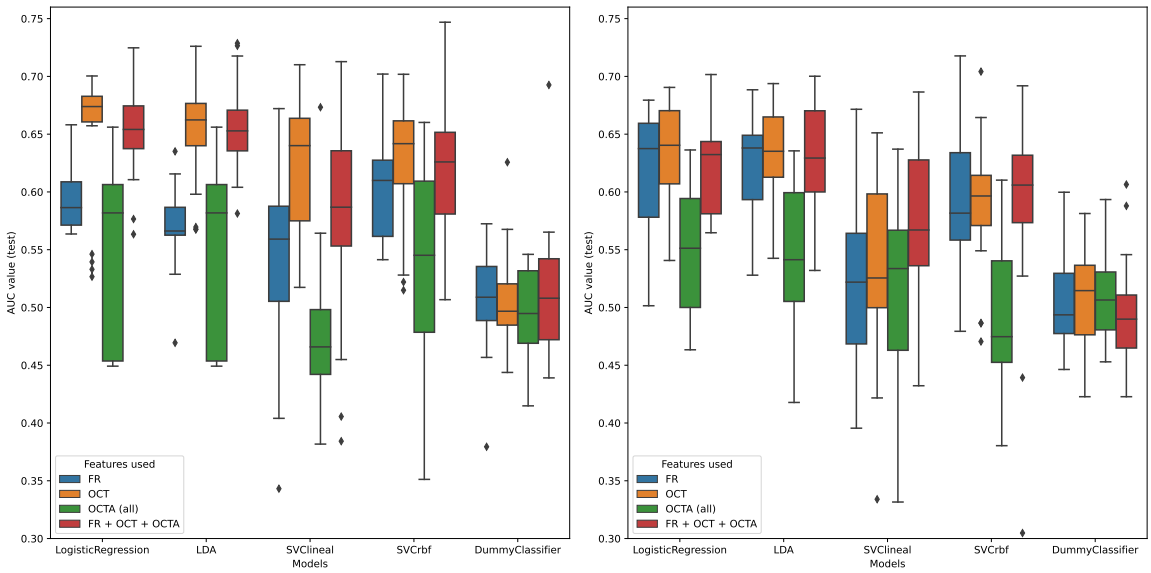


Figure 5.27: Resulting boxplot when discriminating for DR

We can see that OCT features are the ones with best results. The FR and OCTA features yield more or less similar results. The best results are obtained for logistic regression and LDA. We think the reason the SVM classifiers works worse is because the hyper-parameter search was not exhaustive enough, but this should be further investigated.

Discriminating DM

The binary classification task of discriminating if the eye is from a diabetic or not (class 0 *vs* the others). We name this classification task 0-12. We obtain the results shown at the left plot in Figure 5.28. We can see that, again, the OCT features produce the best performance. The FR features work better than random, while the OCTA features are all over the place. Again, we see that the SVM has worse results than the simple classifiers, which further supports the idea of lack of fine tuning of the hyper-parameters.



Boxplot of task 0-12 Boxplot of task 0-12 (without bias)

Figure 5.28: Boxplots of the results when discriminating DM from controls

At this point, we inspected the sources which gave the best results and we found a bias in the data. Specifically, we found that the range of images from 388 to 420 have OCT scans with different noise and level of gray. In Figure 5.29, we can observe some examples.

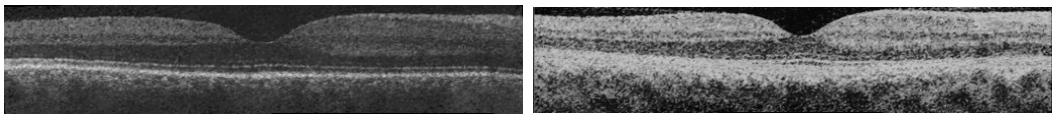


Figure 5.29: Example of the possible bias in the data

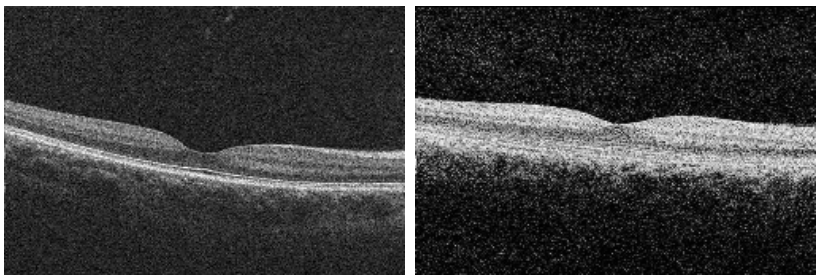


Figure 5.30: Example of the possible bias on the original images

This, in itself, would not be necessarily a problem if it was not because that range of images has more controls than the other classes. In the filtered data, those are 32 individuals of class 0, 4 of class 1 and 1 of class 2. In the non-filtered data those are 55 of class 0, 6 of class 1 and 4 of class 2.

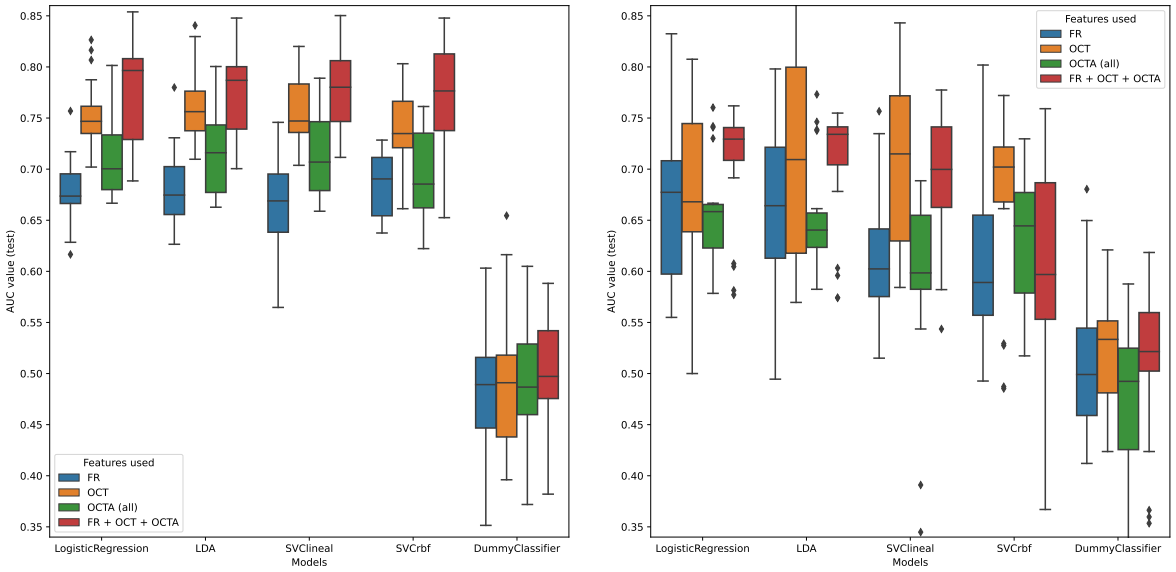
According to the expert, that could be because the lens of the camera equipment was dirty when those images were taken, or something to do with how they were exported from the camera equipment software. In order to progress with this work, we decided to test how the model performs removing those instances. This change means going from having 136 class 0 eyes to have to have 104.

If we re-execute the models after this change we obtain the results shown on the right plot of Figure 5.28. The results of the OCT worsened a bit and have more variance. Oddly enough, the retinography images improved a bit. Since there is no quality filter for the FR images, it could be that the removed instances were difficult instances where the models failed previously. Also, we notice that although the OCTA results worsened a bit on average, they stabilized, exhibiting less variance.

As an extra (improvised) task, we also tried to classify class 0 from class 1 when removing the allegedly easier cases of classes 2,3,4,5. We name this task 0-1. The results are in fact a bit worse, with the mean being around 0.65 AUC with bias and 0.60 without for the best performing model (LR with OCT features). The results are still better than the dummy classifier in both cases, albeit only by a small margin.

Discriminating controls from DR

For completeness, we perform the classification task class 0 versus [2,3,4,5]. Hence, we remove the class 1 which is the majority class. This way, we will emphasize the importance of the bias of the class 0. We call this classification task 0-2. The reason to not have done a three class classification problem is because the problem was originally defined as two binary tasks. Also, given the found bias in the aforementioned range of images, we execute the models with and without the bias. The results can be seen in Figure 5.31.



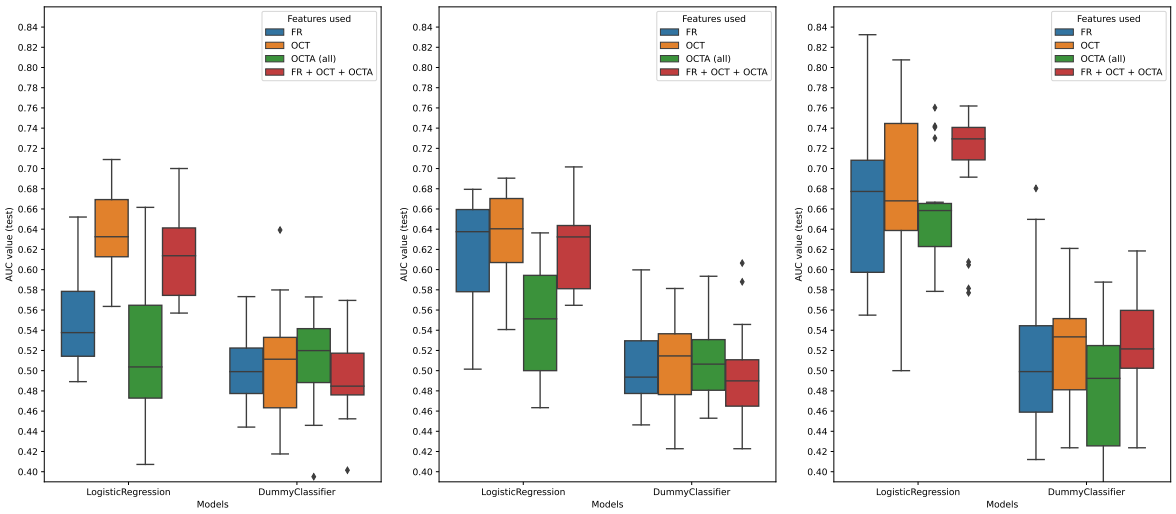
Boxplot of task 0-2 | Boxplot of task 0-2 (without bias)

Figure 5.31: Boxplots of the results when separating controls from DR

It is clear that the results are consistently better and with less variability when including the range with the bias. In fact, in this case all the models perform similarly, even the SVM which gave worse results in the other tasks. Removing the biased range decreases the performance and increases the variability of the results, but it is still better than the dummy classifier. A noticeable detail is that in this classification task, using all the features (FR, OCT, OCTA) gives consistently the best results with and without bias.

Summary

We provide a summary plot for each main classification task side by side, showing only the logistic regression model and dummy classifier results (Figure 5.32).



Boxplot 1-2 Boxplot 0-12 (no bias) Boxplot 0-2 (no bias)

Figure 5.32: Boxplots of the classification tasks side by side

The easiest problem and the one that performs the best is the task 0-2. As for the results of the other two problems 1-2 and 0-12, they are pretty much the same. We also tried an improvised modification of the task 0-12 which we named task 0-1 which is a harder problem, and as expected, the performance was indeed worse. Overall it seems like for some reason, the class 0 is easier to discriminate than it should even after taking the found bias into account. The results of the problem 1-2 are not impressive in comparison to other works in the state of the art. In the other classification tasks though, we obtain promising results which warrant further exploration.

Chapter 6

Conclusions

The results, specially for the OCT images, are promising. Furthermore, by using interpretable models we were able to discover a bias in the data and correct for it. Even after taking the bias into account, promising results were obtained which merit further exploration.

The preprocessing of the images was thorough and satisfactory, and the models learned a parts-based representation. With those, we were able to get results that are fairly comparable to the radiomics approach.

As a overall trend though, we could see that the FR and specially OCTA images lagged behind. The most likely reason being that the vessels exhibit too much variability for a linear factorization method. Manual inspection of the sources supports this hypothesis. Therefore, for those type of images, we conclude that factorization models may not be the best choice and thus other approaches should be explored.

As for the set of objectives of this thesis, we believe to have achieved all of them successfully, albeit with varying levels of success.

- We reviewed the imaging techniques and the NMF literature.
- We explored the retinal image dataset and fixed several data problems.
- We filtered the dataset based on a previous study and implemented the needed preprocessing for the unsupervised factorization models to work.
- We implemented and applied a feature selection and double CV scheme to robustly estimate the performance of the resulting decompositions.

Nevertheless, there are some limitations that must be acknowledged. Regarding the classifier training, there was a possible lack of exhaustive fine-tuning of the SVM classifiers. Also, due to the limited amount of data, we were no able to do a double

CV with more folds. Besides, and perhaps more importantly, even if the results are promising for the given dataset, the unknown that remains is how reproducible would be this approach for other datasets of Type 1 DM.

6.1 Some extensions: Neural Network

Based in a previous work [84], we exploratorily tried to apply transfer learning to the OCTA $6 \times 6mm$ *superficial* images. Authors in that study show good results using transfer learning on a VGG-16 neural network architecture with *ImageNet* weights. They freeze the neural network layer weights except for the last network layers. We tried to reproduce this setting using the same number of frozen network layers and several others. However, we were not able to obtain any relevant results in our dataset. This could be due to many reasons, including on how the implementation of the transfer learning was done, or because of the dataset.

6.2 Future work

Future work could include a non-linear variant of NMF such as kernelized NMF. If non-linear encodings are pursued, perhaps an auto-encoder may be also a good idea, especially if the model explainability is addressed with, for instance, a SHAP (SHapley Additive exPlanations) approach.

Needless to say, *ad-hoc* defined features based on the medical knowledge of the retinal images might lead to improved results.

The limited number of instances could be mitigated by using transfer learning scheme, which would enable the use of deep neural networks, or even deep NMF. Still, if this approach is to be pursued, as shown in the literature, precautions would need to be in place to avoid overfitting.

Also, (semi-)supervised matrix factorization could be used to improve the results. Nevertheless, we would only expect a slight improvement over the non-supervised models. One example would be Discriminant NMF. In this direction, another interesting approach could be factorization machines.

6.3 Acknowledgments

I would like to thank my thesis advisor Alfredo Vellido Alcacena for all the valuable feedback provided during the course of the thesis. Also I would like to thank the

ophthalmologist expert Javier Zarranz-Ventura; without him it would not have been possible the development of this thesis, both for his expertise and for having provided access to the retinal dataset.

I would also like to thank my family for all they have done for me and my friends for being there whenever I was in need.

Bibliography

[1] National Eye Institute. (2019). Diabetic retinopathy. Retrieved May 20, 2021, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy

[2] World Health Organization. (2013). *Diabetes*. Retrieved June 19, 2021, from https://web.archive.org/web/20130826174444/http://www.who.int/mediacentre/factsheets/fs312/en/

[3] National Institute of Diabetes and Digestive and Kidney Diseases. (2014). *Causes of diabetes*. Retrieved June 21, 2021, from https://web.archive.org/web/20160810063435/https://www.niddk.nih.gov/health-information/diabetes/causes

[4] Daneman, D. (2006). Type 1 diabetes. *The Lancet*, *367*(9513), 847–858.

[5] Chiang, J. L., Kirkman, M. S., Laffel, L. M., & Peters, A. L. (2014). Type 1 diabetes through the life span: A position statement of the american diabetes association. *Diabetes care*, *37*(7), 2034–2054.

[6] National Health Service UK. (2018). *Diabetic retinopathy*. Retrieved May 20, 2021, from https://www.nhs.uk/conditions/diabetic-retinopathy/

[7] Grzybowski, A., Brona, P., Lim, G., Ruamviboonsuk, P., Tan, G. S., Abramoff, M., & Ting, D. S. (2020). Artificial intelligence for diabetic retinopathy screening: A review. *Eye*, *34*(3), 451–460.

[8] Prasanna, P., Bobba, V., Figueiredo, N., Sevgi, D. D., Lu, C., Braman, N., Alilou, M., Sharma, S., Srivastava, S. K., Madabhushi, A. et al. (2020). Radiomics-based assessment of ultra-widefield leakage patterns and vessel network architecture in the permeate study: Insights into treatment durability. *British Journal of Ophthalmology*.

[9] Fenner, B. J., Wong, R. L., Lam, W.-C., Tan, G. S., & Cheung, G. C. (2018). Advances in retinal imaging and applications in diabetic retinopathy screening: A review. *Ophthalmology and therapy*, *7*(2), 333–346.

[10] Krawitz, B. D., Mo, S., Geyman, L. S., Agemy, S. A., Scripsema, N. K., Garcia, P. M., Chui, T. Y., & Rosen, R. B. (2017). Acircularity index and axis ratio of the foveal avascular zone in diabetic eyes and healthy controls measured by optical coherence tomography angiography. *Vision research*, *139*, 177–186.

[11] Samara, W. A., Shahlaee, A., Adam, M. K., Khan, M. A., Chiang, A., Maguire, J. I., Hsu, J., & Ho, A. C. (2017). Quantification of diabetic macular ischemia using optical coherence tomography angiography and its relationship with visual acuity. *Ophthalmology*, *124*(2), 235–244.

[12] Werner, J. U., Böhm, F., Lang, G. E., Dreyhaupt, J., Lang, G. K., & Enders, C. (2019). Comparison of foveal avascular zone between optical coherence tomography angiography and fluorescein angiography in patients with retinal vein occlusion. *PLoS ONE*, *14*(6), e0217849.

[13] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), 2402–2410.

[14] Lynch, S. K., Shah, A., Folk, J. C., Wu, X., & Abramoff, M. D. (2017). Catastrophic failure in image-based convolutional neural network algorithms for detecting diabetic retinopathy. *Investigative Ophthalmology & Visual Science*, *58*(8), 3776–3776.

[15] Zarranz-Ventura, J., González, R. A., Fernández, M. I., Medina, J. L., de Viteri Vazquez, M. S., Zapata, M. A., Almuiña, P., Puyuelo, J. A., Pérez, C. B., Bernal-Morales, C., López, J. D., Zamora, J. G., Callén, C. I., de Moura, J., Novo, J., Ortega, M., Penedo, M. G., Bello, J. J. R., & Martín, J. N. R. (2020). *Inteligencia artificial en retina. monografía*. Sociedad Española de Retina y Vítreo.

[16] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D. et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, *24*(9), 1342–1350.

[17] California Healthcare Foundation, & EyePACS. (2015). *Diabetic retinopathy detection*. Retrieved March 31, 2021, from https://www.kaggle.com/c/diabetic-retinopathy-detection

[18] Asia Pacific Tele-Ophthalmology Society. (2019). *Aptos 2019 blindness detection*. Retrieved May 15, 2021, from https://www.kaggle.com/c/aptos2019-blindness-detection

[19] Xu, G. (2019). *1st place solution summary*. Retrieved May 21, 2021, from https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/108065

[20] Yu, Y., & Zhu, H. (2021). Retinal vessel segmentation with constrained-based nonnegative matrix factorization and 3d modified attention u-net. *EURASIP Journal on Image and Video Processing*, *2021*(1), 1–21.

[21] Núñez, L. M., Romero, E., Julià-Sapé, M., Ledesma-Carbayo, M. J., Santos, A., Arús, C., Candiota, A. P., & Vellido, A. (2020). Unraveling response to temozolomide in preclinical gl261 glioblastoma with mri/mrsi using radiomics and signal source extraction. *Scientific reports*, *10*(1), 1–13.

[22] Rhcastilhos, J. (2007). *Diagram of the human eye in english. it shows the lower part of the right eye after a central and horizontal section*. Retrieved June 11, 2021, from https://commons.wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg

[23] Ophthalmology Physicians & Surgeons. (2021). *Treating diabetic retinopathy*. Retrieved April 4, 2021, from https://www.eyeops.com/contents/our-services/eye-diseases/diabetic-retinopathy

[24] Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., & Masis, M. (2013). Classification of diabetic retinopathy and diabetic macular edema. *World journal of diabetes*, *4*(6), 290.

[25] Taylor, D. (2017). Diabetic eye screening programme grading definitions for referable disease public health england leads the nhs screening programmes.

[26] Voets, M., Møllersen, K., & Bongo, L. A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, *14*(6), e0217541.

[27] Basit, A., & Egerton, S. (2013). Bio-medical imaging: Localization of main structures in retinal fundus images. *IOP Conference Series: Materials Science and Engineering*, *51*(1), 012009.

[28] Saine, P., & Tyler, M. (2002). *Ophthalmic photography: Retinal photography, angiography, and electronic imaging*. Butterworth-Heinemann.

[29] Stolte, S., & Fang, R. (2020). A survey on medical image analysis in diabetic retinopathy. *Medical image analysis*, *64*, 101742.

[30] Scholl, H. (2018). *The research work of professor hendrik scholl*. Retrieved March 13, 2021, from http://www.vision-research.eu/index.php?id=1169

[31] Wikipedia. (2021a). *Tomography*. Retrieved June 11, 2021, from https://en.wikipedia.org/wiki/Tomography

[32] George Wildeman, L. M., Hafeez Dhalla. (2016). *What is oct and how can it help ophthalmologists acquire high resolution information on ocular tissue?* Retrieved June 8, 2021, from https://www.leica-microsystems.com/science-lab/what-is-oct-and-how-can-it-help-ophthalmologists-acquire-high-resolution-information-on-ocular-tissue/

[33] Wikipedia. (2021b). *Optical coherence tomography*. Retrieved June 11, 2021, from https://en.wikipedia.org/wiki/Optical_coherence_tomography

[34] Rogers, J., Podoleanu, A., Dobre, G., Jackson, D., & Fitzke, F. (2001). Topography and volume measurements of the optic nerve using en-face optical coherence tomography. *Optics express*, *9*, 533–45. https://doi.org/10.1364/OE.9.000533

[35] Barraso, M., Alé-Chilet, A., Hernández, T., Oliva, C., Vinagre, I., Ortega, E., Figueras-Roca, M., Sala-Puigdollers, A., Esquinas, C., Esmatjes, E. et al. (2020). Optical coherence tomography angiography in type 1 diabetes mellitus. report 1: Diabetic retinopathy. *Translational vision science & technology*, *9*(10), 34–34.

[36] Zarranz-Ventura, J. (2017). Detección del edema macular diabético en atención primaria con tomografía óptica de coherencia.

[37] Gillis, N. (2014). The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 257–291.

[38] Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, *5*(9).

[39] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

[40] Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126.

[41] Wang, Y.-X., & Zhang, Y.-J. (2012). Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1336–1353.

[42] Vavasis, S. A. (2010). On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, *20*(3), 1364–1377.

[43] Vasiloglou, N., Gray, A. G., & Anderson, D. V. (2009). Non-negative matrix factorization, convexity and isometry. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 673–684.

[44] Klingenberg, B., Curry, J., & Dougherty, A. (2009). Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition*, *42*(5), 918–928.

[45] Donoho, D., & Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? *Proceedings of the 16th International Conference on Neural Information Processing Systems*, 1141–1148.

[46] Gillis, N. (2020). *Nonnegative matrix factorization*. SIAM.

[47] Kim, H., & Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, *23*(12), 1495–1502.

[48] Kim, J., He, Y., & Park, H. (2014). Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, *58*(2), 285–319.

[49] Lee, D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems*. MIT Press.

[50] Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural computation*, *23*(9), 2421–2456.

[51] Chu, M., Diele, F., Plemmons, R., & Ragni, S. (2004). Optimality, computation, and interpretation of nonnegative matrix factorizations. *SIAM Journal on Matrix Analysis*.

[52] Gonzalez, E. F., & Zhang, Y. (2005). *Accelerating the lee-seung algorithm for nonnegative matrix factorization* (tech. rep.).

[53] Cichocki, A., & Phan, A.-H. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, *92*(3), 708–721.

[54] Pikna, M. (2019). Evaluating performance of an image compression scheme based on non-negative matrix factorization.

[55] Kanagal, B., & Sindhwani, V. (2010). Rank selection in low-rank matrix approximations: A study of cross-validation for nmfs. *Proc Conf Adv Neural Inf Process*, *1*, 10–15.

[56] Gillis, N. (2019). *Linear dimensionality reduction for data analysis*. Retrieved June 17, 2021, from https://www.youtube.com/watch?v=HNvVKVPwB7s

[57] Lee, S. (2020). Estimating the rank of a nonnegative matrix factorization model for automatic music transcription based on stein's unbiased risk estimator. *Applied Sciences*, *10*(8), 2911.

[58] Boutsidis, C., & Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, *41*(4), 1350–1362.

[59] Zitnik, M., & Zupan, B. (2012). Nimfa: A python library for nonnegative matrix factorization. *Journal of Machine Learning Research*, *13*, 849–853.

[60] Albright, R., Cox, J., Duling, D., Langville, A. N., & Meyer, C. (2006). *Algorithms, initializations, and convergence for the nonnegative matrix factorization* (tech. rep.). Tech. rep. 919. NCSU Technical Report Math 81706.

[61] Scikit-learn developers. (n.d.[a]). *Nmf with a beta-divergence*. Retrieved June 12, 2021, from https://scikit-learn.org/stable/modules/decomposition.html#nmf-with-a-beta-divergence

[62] Gaussier, E., & Goutte, C. (2005). Relation between plsa and nmf and implications. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 601–602.

[63] Ding, C., Li, T., & Peng, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. *AAAI*, *42*, 137–43.

[64] Ding, C. H., Li, T., & Jordan, M. I. (2008). Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, *32*(1), 45–55.

[65] Gillis, N., & Vavasis, S. A. (2013). Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, *36*(4), 698–714.

[66] Berry, M. W., Gillis, N., & Glineur, F. (2009). Document classification using nonnegative matrix factorization and underapproximation. *2009 IEEE International Symposium on Circuits and Systems*, 2782–2785.

[67] Gillis, N., & Glineur, F. (2010). Using underapproximations for sparse nonnegative matrix factorization. *Pattern recognition*, *43*(4), 1676–1687.

[68] Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural networks*, *13*(4-5), 411–430.

[69] Scikit-learn developers. (n.d.[b]). *Decomposing signals in components (matrix factorization problems)*. Retrieved June 17, 2021, from https://scikit-learn.org/0.24/modules/decomposition.html

[70] Gillis, N. (2017). Introduction to nonnegative matrix factorization. *arXiv preprint arXiv:1703.00663*.

[71] Li, T., & Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. *Sixth International Conference on Data Mining (ICDM'06)*, 362–371.

[72] Ding, C., He, X., & Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. *Proceedings of the 2005 SIAM international conference on data mining*, 606–610.

[73] Chabacano. (2008). *Diagram showing overfitting of a classifier*. Retrieved June 17, 2021, from https://en.wikipedia.org/wiki/File:Overfitting.svg

[74] Scikit-learn developers. (n.d.[c]). *Cross-validation: Evaluating estimator performance*. Retrieved June 4, 2021, from https://scikit-learn.org/0.24/modules/cross_validation.html

[75] OpenCV contributors. (2021). *Image processing in opencv*. Retrieved April 13, 2021, from https://docs.opencv.org/4.5.2/d2/d96/tutorial_py_table_of_contents_imgproc.html

[76] Huamán, A. (n.d.). *Changing the contrast and brightness of an image*. Retrieved June 21, 2021, from https://docs.opencv.org/4.5.2/d3/dc1/tutorial_basic_linear_transform.html

[77] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[78] Zarranz-Ventura, J., Barraso, M., Alé-Chilet, A., Hernandez, T., Oliva, C., Gascón, J., Sala-Puigdollers, A., Figueras-Roca, M., Vinagre, I., Ortega, E. et al. (2019). Evaluation of microvascular changes in the perifoveal vascular network using optical coherence tomography angiography (octa) in type i diabetes mellitus: A large scale prospective trial. *BMC medical imaging*, *19*(1), 1–6.

[79] Junior, S. B., & Welfer, D. (2013). Automatic detection of microaneurysms and hemorrhages in color eye fundus images. *International Journal of Computer Science & Information Technology*, *5*(5), 21.

[80] Graham, B. (2015). *Kaggle diabetic retinopathy detection competition report*. Retrieved March 31, 2021, from https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15801

[81] Graham, B. (2014). Fractional max-pooling. *arXiv preprint arXiv:1412.6071*.

[82] Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ*, *2*, e453.

[83] Wikipedia. (n.d.). *Mutual information*. Retrieved June 4, 2021, from https://en.wikipedia.org/wiki/Mutual_information

[84] Le, D., Alam, M., Yao, C. K., Lim, J. I., Hsieh, Y.-T., Chan, R. V., Toslak, D., & Yao, X. (2020). Transfer learning for automated octa detection of diabetic retinopathy. *Translational Vision Science & Technology*, *9*(2), 35–35.