UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

BE-OPTICAL
Advanced Biomedical Optical
Imaging and Data Analysis

Machine learning methods for the characterization and classification of complex data

Author:

Pablo Amil Marletti

MSc in Engineering physics

Supervisor:

Cristina Masoller

PhD in Physics

Thesis submitted in fulfillment of the requirements of the degree of
Doctor in Computational and Applied Physics

Universitat Politècnica de Catalunya

Machine learning methods for the characterization and classification of complex data

A thesis submitted by

Pablo Amil Marletti

MSc in Engineering physics

in fulfillment of the requirements of the degree of
Doctor in Computational and Applied Physics

Supervisor:
Cristina Masoller

PhD in Physics

Departament de Física
Terrassa, December 2019

Abstract

This thesis work presents novel methods for the analysis and classification of medical images and, more generally, complex data. First, an unsupervised machine learning method is proposed to order anterior chamber OCT (Optical Coherence Tomography) images according to a patient's risk of developing angle-closure glaucoma. In a second study, two outlier finding techniques are proposed to improve the results of above mentioned machine learning algorithm, we also show that they are applicable to a wide variety of data, including fraud detection in credit card transactions. In a third study, the topology of the vascular network of the retina, considering it a complex tree-like network is analyzed and we show that structural differences reveal the presence of glaucoma and diabetic retinopathy. In a fourth study we use a model of a laser with optical injection that presents extreme events in its intensity time-series to evaluate machine learning methods to forecast such extreme events.

Resum

Mètodes d'aprenentatge automàtic per a la caracterització i classificació de dades complexes

Aquesta tesi desenvolupa nous mètodes per a l'anàlisi i la classificació d'imatges mèdiques i dades complexes. Hem proposat, primer, un mètode d'aprenentatge automàtic sense supervisió que ordena imatges OCT (tomografia de coherència òptica) de la cambra anterior de l'ull en funció del grau de risc del pacient de patir glaucoma d'angle tancat. Després, hem desenvolupat dos mètodes de detecció automàtica d'anomalies que hem utilitzat per millorar els resultats de l'algoritme anterior, però que la seva aplicabilitat va molt més enllà, sent útil, fins i tot, per a la detecció automàtica de fraus en transaccions de targetes de crèdit. Mostrem també, com en analitzar la topologia de la xarxa vascular de la retina considerant-la una xarxa complexa, podem detectar la presència de glaucoma i de retinopatia diabètica a través de diferències estructurals. Finalment, hem estudiat un làser amb injecció òptica, el qual presenta esdeveniments extrems en la sèrie temporal d'intensitat. Hem avaluat diferents mètodes per tal de predir-los.

Resumen

Métodos de aprendizaje automático para la caracterización y clasificación de datos complejos

El presente trabajo de tesis desarrolla nuevos métodos para el análisis y clasificación de imágenes médicas y datos complejos en general. Primero, proponemos un método de aprendizaje automático sin supervisión que ordena imágenes OCT (tomografía de coherencia óptica) de la cámara anterior del ojo en función del grado de riesgo del paciente de padecer glaucoma de ángulo cerrado. Luego, desarrollamos dos métodos de detección automática de anomalías que utilizamos para mejorar los resultados del algoritmo anterior, pero que su aplicabilidad va mucho más allá, siendo útil, incluso, para la detección automática de fraudes en transacciones de tarjetas de crédito. Mostramos también, cómo al analizar la topología de la red vascular de la retina considerándola una red compleja, podemos detectar la presencia de glaucoma y de retinopatía diabética a través de diferencias estructurales. Estudiamos también un modelo de un láser con inyección óptica que presenta eventos extremos en la serie temporal de intensidad para evaluar diferentes métodos de aprendizaje automático para predecir dichos eventos extremos.

Agradecimientos

Realizar un doctorado es un trabajo que mezcla momentos de gran satisfacción con momentos de frustración absoluta. Si bien dicho trabajo es un trabajo de investigación individual, este no es posible sin la ayuda y el apoyo de un montón de gente dentro, y fuera, del ámbito académico. Quiero tomarme unas líneas para agradecer a todos los que me brindaron su apoyo en algún punto del camino.

A mi supervisora y compatriota, Cristina Masoller, que me invitó a realizar este trabajo, me aconsejó y me guio durante todo el proceso. A Ulrich Parlitz, quien me recibió durante mis estadías en el instituto Max Plank en Göttingen y con quien colaboramos intensamente en el primer tramo de mi trabajo. Al equipo médico y técnico del Instituto de Microcirugía ocular en Barcelona, en especial a las doctoras Elena Arrondo y Cecilia Salinas que dedicaron su tiempo a ayudarme a interpretar las imágenes, y a Laura González quien me asistió en la recopilación de imágenes y revisión de las historias clínicas. A Irene Sendiña, a Antonio Pons, a Fabián Reyes, a Nahuel Almeira, y a Miguel C. Soriano.

También debo agradecer a las personas que me aguantaron y me apoyaron moralmente, a Donatus Halpaap, a Dario Zappala, a Maria Masoliver, a Hossam Selim, a Sandeep Gawali, a Jordi Tiana, a Carlos Quintero, y muy especialmente a mi esposa Lina Huertas y a mi familia.

Quiero reconocer el apoyo económico que he recibido de la Comisión Europea, a través del proyecto Be-Optical del programa H2020 y de la Universitat Politècnica de Catalunya.

Contents

List of acronyms

AS-OCT	Anterior Segment Optical Coherence Tomography
AUC	Area Under the receiver operating characteristic Curve
C-NDD	Central Node Distance Distribution
CMWD	Central Mean Weight Distribution
COV	Covariance Matrix
d2CM	Distance to the Center-of-Mass
DR	Diabetic Retinopathy
FD	Fractal Dimension
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GCC	Giant Connected Component
HOG	Histogram of Oriented Gradients
IMO	Instituto de Microcirugía Ocular (Ocular Microsurgery Institute)
IsoMap	Isometric Feature Mapping
JS	Jensen-Shanon
KNN	k-Nearest Neighbours
MAPE	Mean Absolute Percentage Error
MARE	Mean Absolute Relative Error
MDS	Multidimensional Scaling
NDD	Node Distance Distribution
NLDR	Nonlinear Dimensionality Reduction
NN	Neural Network
OCSVM	One Class Support Vector Machine

OCT	Optical Coherence Tomography
OS	Outlier Score
PCA	Principal Component Analysis
PDF	Probability Distribution Function
RC	Reservoir Computing
ROC	Receiver Operating Characteristic
SSE	Sum of Squared Errors
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
WDD	Weighted Degree Distribution

Chapter 1.

Introduction

1.1. Background and motivation

Providing high quality health care to an ever increasing, aging population is now, more than a ever, a difficult challenge that society has to face.

Virtually every health center now-a-days is generating enormous amounts of data. Most imaging techniques and tests are digitally stored, making it possible to analyze using big data techniques. The use of big data techniques is increasingly popular in the clinic, but a lot of research and development is needed to deploy its full capabilities. Two main challenges have to be addressed in order to make an efficient use of the data, namely availability and preparation of the data and algorithmic use of such data. In other words, all the data needs to be accessible and in a format that it can be decoded by the central processing algorithm. Also, new algorithms to process and to draw useful conclusions using large amounts of data need to be developed specifically for medical data. In this thesis, we aim to provide methods and algorithms to progress towards the second challenge, i.e., we aim at developing algorithms that can, unsupervisedly, extract relevant information from complex datasets.

The analysis of complex data in all areas of science is usually a challenging task. Complex data is the result of the measurement of a complex system (such as the human body, or a chaotic laser), whose patterns are not trivial to discover in an algorithmic manner. The human brain is very good at recognizing patterns and rapidly learns how to draw conclusions. Artificial intelligence is a research area whose main purpose is to mimic the function of the brain in an algorithmic manner. The use of artificial intelligence techniques with medical data has provided many useful results. For example, in Ref. [1] a method for diagnosing patients with congestive heart failure is shown to give excellent results, in Ref. [2] the authors propose a platform for analyzing ocular images to diagnose congenital cataracts in patients, in Ref. [3] optical coherent tomography (OCT) images are analyzed to diagnose two different ocular diseases.

The use of such techniques can help doctors to optimize their time, making faster and more accurate diagnosis, which translates to improved and lower-cost health care to the population.

Therefore, this thesis is focused on the development of new artificial intelligence techniques, in particular using machine learning approaches, to analyze complex data, specifically ophthalmic images. In the process we aim to learn and develop new techniques that can be useful in other areas of science, such as outlier detection and extreme event prediction.

1.2. Machine learning techniques

Machine learning [4–6] is an area of artificial intelligence which relies on statistical models to identify patterns in the data. We used machine learning as the main tool throughout this thesis, in chapter 2 (Ref. [7]) we used unsupervised manifold learning algorithms to obtain a meaningful ordering of the OCT images (data is described in section 1.4.1) in a plane. In chapter 3 (Ref. [8]) we developed two unsupervised machine learning algorithms to detect outliers (entries that differ from the normality) in databases. In chapter 4 (Ref. [9]), we used unsupervised dimensionality reduction techniques to differentiate eye diseases. In chapter 5 (Ref. [10]) we used several (supervised) machine learning tools to forecast the evolution of a chaotic laser.

Most machine learning techniques can be formalized as a function approximation problem [4] in which for each data point, $\boldsymbol{x}$, in a given space, $\mathcal{X}$, there exists a point, y, in an image space $\mathcal{Y}$. Such function

$$f : \mathcal{X} \to \mathcal{Y}; y = f\left(\boldsymbol{x}\right), \tag{1.2.1}$$

is the objective function, and different algorithms compute approximations, $\tilde{f}$ (the *learned function*), of such function f. A more general approach to the formalization of machine learning proposed in Ref. [5] is determining properties of the conditional probability $P\left(y|\boldsymbol{x}\right)$.

We can classify machine learning algorithms according to the data that is used for training and the type of function we want to *learn* as supervised learning or unsupervised learning.

1.2.1. Supervised learning

This class of algorithms take examples of inputs and outputs of the objective function, f, and use them to generate an approximation $\tilde{f}$. The way these algorithms work is by defining a function search space, $\mathcal{F}$, from which to choose $\tilde{f}$, and a searching algorithm that usually minimizes an error measure between $\tilde{f}$ and f. We can further classify algorithms according to the task they perform (the type of set $\mathcal{Y}$ is) as *regression learning* when $\mathcal{Y}$ is some power of the real set ($\mathbb{R}^n$) and *supervised classification* algorithms when $\mathcal{Y}$ is a categorical set (i.e., a set with a finite number of symbols).

Under this paradigm, the well-known linear regression algorithm can be seen as a supervised regression learning algorithm, where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and

$$\mathcal{F} = \left\{ \tilde{f}\left(\boldsymbol{x}\right) = a\boldsymbol{x} + b \,|\, (a,b) \in \mathbb{R}^2 \right\}.$$

The searching algorithm chooses the pair (a, b) that defines $\tilde{f}$ by minimizing the sum of squared errors

$$\text{SSE} = \sum_i \left(\tilde{f}\left(\boldsymbol{x}_i\right) - y_i \right)^2,$$

where $y_i = f\left(\boldsymbol{x}_i\right)$ are examples of the objective function.

Another example of supervised learning is the k-nearest neighbors algorithm, which we used in chapter 5 (Ref. [10]). It defines the learned function as

$$\tilde{f}\left(\boldsymbol{x}\right) = \frac{1}{k} \sum_{i \in \mathcal{N}} y_i,$$

where $\mathscr{N}$ is the neighborhood of $\boldsymbol{x}$ defined as the set of k indexes of the examples whose distance $\operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}_i)$ are smallest.

1.2.2. Unsupervised learning

This class of algorithms use only examples of the domain set ($\mathcal{X}$) of the objective function. They can be used for nonlinear dimensionality reduction, when $\mathcal{Y}$ is some power of the real set ($\mathbb{R}^n$), for clustering when $\mathcal{Y}$ is a categorical set, or for anomaly detection.

A widely used algorithm for performing dimensionality reduction is the Principal Component Analysis (PCA). PCA is a linear method that finds a set of orthogonal axis to transform the data into a new representation. In this new representation, each coordinate of the data is (linearly) uncorrelated with all the others. If one orders these coordinates in decreasing standard deviation, it is possible to maintain only a few dimensions with minimal information loss.

In this thesis we have extensively used the IsoMap [11] algorithm which is a nonlinear dimensionality reduction technique or manifold learning technique, that, basically, extends the PCA idea to nonlinear transformations. It aims to translate high-dimensional data into a lower-dimensional representation that preserves the structure of the original data. The IsoMap algorithm starts by computing the pair-wise distances of the original data, obtaining a distance matrix

$$D_{ij} = \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)\,.$$

It follows by constructing a graph maintaining local distances, this can be achieved by either a) disregarding all the distances longer than a certain threshold, or b) finding the k nearest neighbors of each point and keeping only the links that corresponds to these nearest neighbors. In this way an adjacency matrix

$$D'_{ij} = \begin{cases} D_{ij} & if \quad i \text{ is nearest neighbor of } j \text{ or vice-versa} \\ 0 & otherwise \end{cases}$$

is obtained. Then, a new distance matrix that contains the length of the shortest paths between each pair of nodes is computed. This may be done using the Dijkstra's algorithm [12] or the Floyd's algorithm [13], ending up with a new matrix, D^G_{ij}, that approximates the geodesic distances. As a last step multidimensional scaling (MDS) [14] is applied to D^G_{ij} obtaining a d-dimensional embedding for the data.

We have also used the t-SNE algorithm [15] which performs a similar task to the IsoMap with the drawbacks that it doesn't preserve local distances and that it is a stochastic method, meaning that the output depends on the seed of the random number generator.

1.2.3. Performance metrics

Several metrics can be used to quantify the performance of machine learning algorithms. The metrics to be used depend on the type of machine learning algorithm and the data available. Here we describe the performance metrics that we have used in this thesis.

1.2.3.1. Confusion matrix

The confusion matrix is a compact way to show the performance of classification algorithms. It shows, for each element in $\mathcal{Y}$, how many items in the test set were correctly classified by $\tilde{f}$, and how many were incorrectly classified. Formally, let $\mathcal{Y} = \{v_1, v_2, ..., v_N\}$, and a test set $y_k = f(\boldsymbol{x}_k)$. Then the confusion matrix, $\mathbf{c} \in \mathbb{R}^{N \times N}$, has the elements

$$c_{i,j} = \# \{k \,|\, \tilde{y}_k = v_i \wedge y_k = v_j\},$$

where $\tilde{y}_k = \tilde{f}(\boldsymbol{x}_k)$.

For example, let us have $\mathcal{Y} = \{a, b, c\}$, a test set, $(\boldsymbol{x}_k, y_k)$, and its predictions $\tilde{y}_k$ as:

k	y_k	$\tilde{y}_k$
1	a	a
2	a	a
3	a	b
4	b	a
5	b	c
6	b	b
7	c	c
8	c	c
9	c	c

then

$$\mathbf{c} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 3 \end{pmatrix}.$$

For binary classification ($\mathcal{Y} = \{1, 0\} = \{"\text{positive}", "\text{negative}"\}$) the entries of the confusion matrix have special names:

- $c_{1,1}$ True positives (TP). The number of elements that were correctly predicted as positive ($y_k = \tilde{y}_k = 1$).
- $c_{2,2}$ True negatives (TN). The number of elements that were correctly predicted as negative ($y_k = \tilde{y}_k = 0$).
- $c_{1,2}$ False positives (FP). The number of elements that were incorrectly predicted as positive ($y_k = 0$ and $y = 1$).
- $c_{2,1}$ False negatives (FN). The number of elements that were incorrectly predicted as negative ($y_k = 1$ and $y = 0$).

Using these metrics, other metrics may be defined, such as:

- Sensitivity (or true positive rate, TPR) as $\frac{c_{1,1}}{c_{1,1}+c_{2,1}}$.
- Specificity (or true negative rate, TNR) as $\frac{c_{2,2}}{c_{1,2}+c_{2,2}}$.

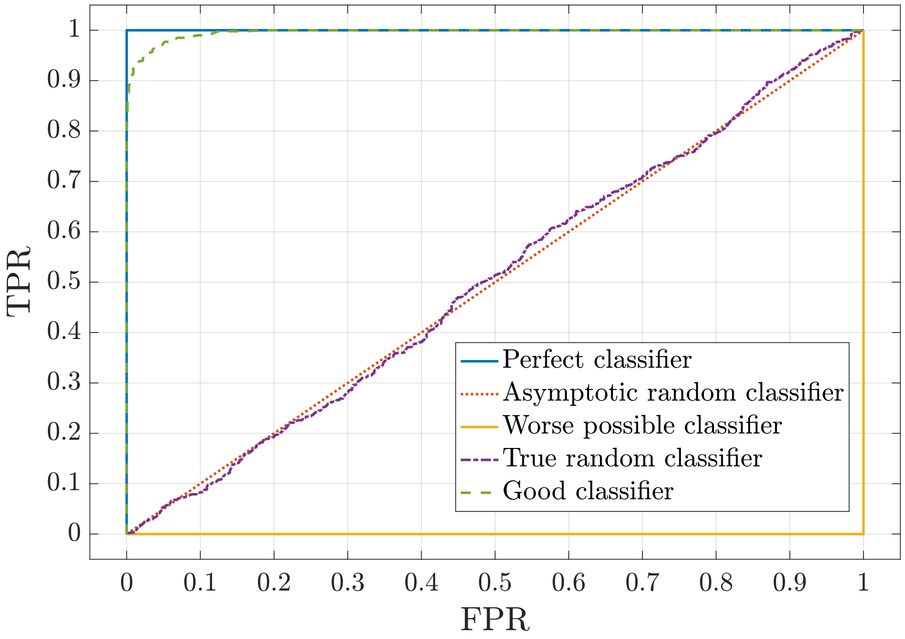


Figure 1.2.1.: Examples of ROC curves for various classifiers.

- Fall-out (or false positive rate FPR) as $\frac{c_{1,2}}{c_{1,2}+c_{2,2}}$.
- Precision as $\frac{c_{1,1}}{c_{1,1}+c_{1,2}}$.

1.2.3.2. Receiver operating characteristic (ROC) curve

This curve is used for assessing the performance of binary classification algorithms, formally when $\mathcal{Y} = \{1, 0\}$ ($\upsilon_1 = 1 =$ "positive", $\upsilon_2 = 0 =$ "negative") or any two-element set. To generate this curve, we vary some internal parameter of the classification algorithm (usually a threshold) and compute the TPR and the FPR.

For each value of the varied parameter, we have the values of the TPR and the FPR, which we can show in a graphical way by plotting the FPR in the x-axis and the TPR in the y-axis. In this representation, a point in the top left corner (FPR=0, TPR=1) corresponds to perfect classification and a point on the bottom right corner (FPR=1, TPR=0) corresponds to wrong classification of every single test point. Therefore, we want our plots to be near the top left corner, see Fig. 1.2.1 for example curves.

1.2.3.3. Area under the ROC curve (AUC)

As we have explained, the ROC curve is a useful visual way for assessing the performance of binary classification algorithms. When varying a threshold or an internal parameter we usually obtain a curve from bottom left corner (FPR=0, TPR=0) to the top right corner (FPR=1, TPR=1). When the varied parameter is a threshold at the end of the algorithm the algorithms classifies all examples in one class when the threshold is $+\infty$ and in the other class if the threshold is $-\infty$. The ROC curve in this case is a monotonically increasing function.

The area under this curve is an indicator of how good the classification algorithm is regardless of the threshold or parameter that is varied. When the AUC $= 1$, the algorithm is able to produce perfect classification for some value of the threshold, in contrast random guessing gives AUC=0.5 when the number of test points tends to infinity.

1.2.3.4. Precision-recall curve

Similarly to the ROC curve, this curve plots two rates as some threshold or internal parameter of the binary classification algorithm is varied. In this case the x-axis is the same as the y-axis in the ROC curve, the TPR also known as *recall*, while the y-axis is the *precision*.

This alternative to the ROC curve is used for very imbalanced scenarios [16] (scenarios in which the data contains a lot more points of one case than the other). In such scenarios, the ROC curve might appear very good for classifiers slightly better than random while the precision-recall curve gives more meaningful results.

1.2.3.5. Average precision

Similarly to the AUC, the average precision is the area under the precision-recall curve. As the x-axis (recall) spans the interval $[0, 1]$, the area under the curve can be interpreted as an average of the precision (which is in the y-axis). We used this metric in chapter 3 (Ref. [8]).

1.2.3.6. p-value

The p-value is a very general concept that, basically, estimates the probability of the null hypothesis being true by computing the probability of the data being true given that the null hypothesis is. When designing a statistical test for the null hypothesis that two groups have the same distribution of a certain feature, it can be used as a method to evaluate how well such feature distinguishes between the two groups.

This can be done by testing the null hypothesis that the two groups come from independent random samples from normal distributions with equal means and equal but unknown variances (known as two-sample t-test). This is the test we used in chapter 2 (Ref. [9]).

Although widely used in science, the p-value is only meaningful in relation to the null hypothesis [17–19] and interpretations have to be carried out carefully.

1.2.3.7. Pearson correlation coefficient

This coefficient is usually used to asses how similar the variability of two vectors of the same length is. It can be used to evaluate the performance of regression learning techniques by comparing the vector of true answers (y_k) and the predicted answers ($\tilde{y}_k$). Formally it is calculated as

$$\rho = \frac{\sum_k (y_k - < y_k >)(\tilde{y}_k - < \tilde{y}_k >)}{\sqrt{\sum_k (y_k - < y_k >)^2}\sqrt{(\tilde{y}_k - < \tilde{y}_k >)^2}}.$$

Its output, ρ, is a number between -1 and 1, where $\rho = 0$ means no correlation, and $|\rho| = 1$ means perfect correlation. The sign of ρ tells whether y_k grows (positive ρ) as $\tilde{y}_k$ or it decreases (negative ρ). In this thesis we have used this measure in chapter 2 (Ref. [7]).

1.2.3.8. Mean absolute relative error (MARE) and mean absolute percentage error (MAPE)

These two closely related measures are useful for evaluating regression learning techniques. The formal definition is

$$\mathrm{MARE} = \frac{1}{K}\sum_{k=1}^{i=K}\frac{|\tilde{y}_k - y_k|}{y_k}$$

for the MARE, where K is the total number of test items.

The MAPE is simply the same quantity as a percentage ($\mathrm{MAPE} = 100 \times \mathrm{MARE}$). In this thesis (chapter 5 and Ref. [10]) we have used the MARE to quantify the accuracy of machine learning prediction algorithms.

1.3. Data analysis and complex networks

In this thesis we have also used several analysis methods based on complex networks. Complex networks are mathematical structures (also known as graphs) made up of two kinds of elements, nodes and links, where a link is a connection between two nodes that can have a weight (the strength of the link) and a directionality. We used networks in one way or another in all our works, but mainly in chapters 3 and 4 (Refs. [8, 9]).

A network of N nodes is defined by its adjacency matrix, $\mathbf{A}$, that is an N-by-N matrix whose entries represent whether a link between nodes i and j exist ($A_{i,j} = 1$) or not ($A_{i,j} = 0$). In directed networks, the adjacency matrix is not symmetric $A_{i,j} \neq A_{j,i}$. For undirected graphs, links have the same strength in both directions and thus the adjacency matrix is symmetric $A_{i,j} = A_{j,i}$.

Networks can be used to study a wide variety of real-world problems [20], in chapter 3 (Ref. [8]) we use networks to represent relationships between items in a database, in this case, each item in a database was represented as a node, and the links connecting them where weighted according to their pair-wise distance. In chapter 4 (Ref. [9]) we studied the vessel network that can be extracted from in eye fundus photographs, we represented bifurcation points and end points of the vessels as nodes, and the vessel sections that connected these nodes as links.

Given a network, we can define a path of length M as sequence p_k ($k = 1, .., M$) of nodes such that $A_{(p_{k-1}),(p_k)} = 1\,\forall k$. An undirected graph is said to be connected if for every pair i, j there exists a path p_k such that $p_1 = i$ and $p_M = j$. If the graph is not connected, it can be divided into connected components or subgraphs. A subgraph is a new graph formed with a subset of the nodes in the original graph and the links that correspond to both nodes in the subset. The connected components are subgraphs that are connected and that there is no link that connects a node of the subgraph to a node outside of it (see an example in Fig. 1.3.1). The concept of connected components is the main idea behind one of the methods proposed in chapter 5 (Ref. [10]).

In unweighted networks it is often interesting to study the so called degree of the nodes. Such degree is simply the number of links that a certain node has

$$\deg(j) = \sum_{i=1}^{i=N} A_{i,j}.$$

As $A_{ij} = 1$ when there is a link and $A_{ij} = 0$ otherwise, the sum equals the degree of the node j. As the degree is a function of the node, studying the distribution of this values is often of interest, the

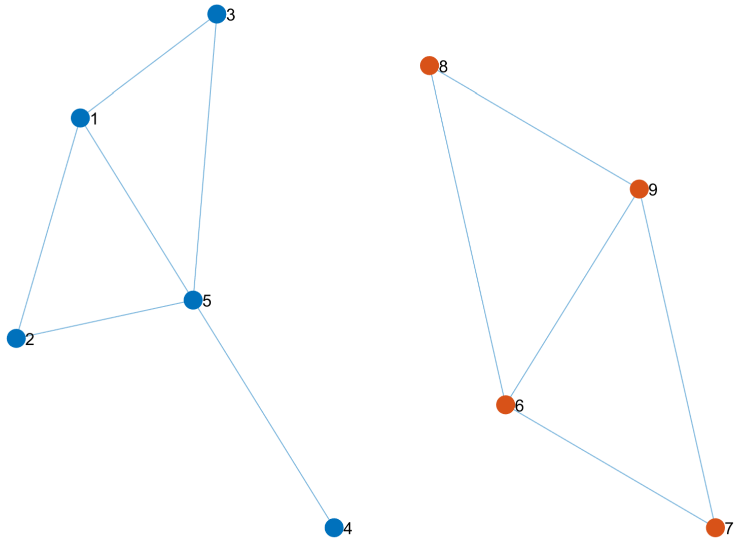


Figure 1.3.1.: Example of a graph with two connected components, one shown in blue (nodes 1 to 5) and another shown in orange (nodes 6 to 9). Graph nodes are represented by filled circles while links are represented by straight lines connecting the corresponding nodes.

so called degree distribution. The degree distribution is defined as

$$P\left(k\right)=\frac{\#\left\{j\,|\,\deg\left(j\right)=k\right\}}{N},$$

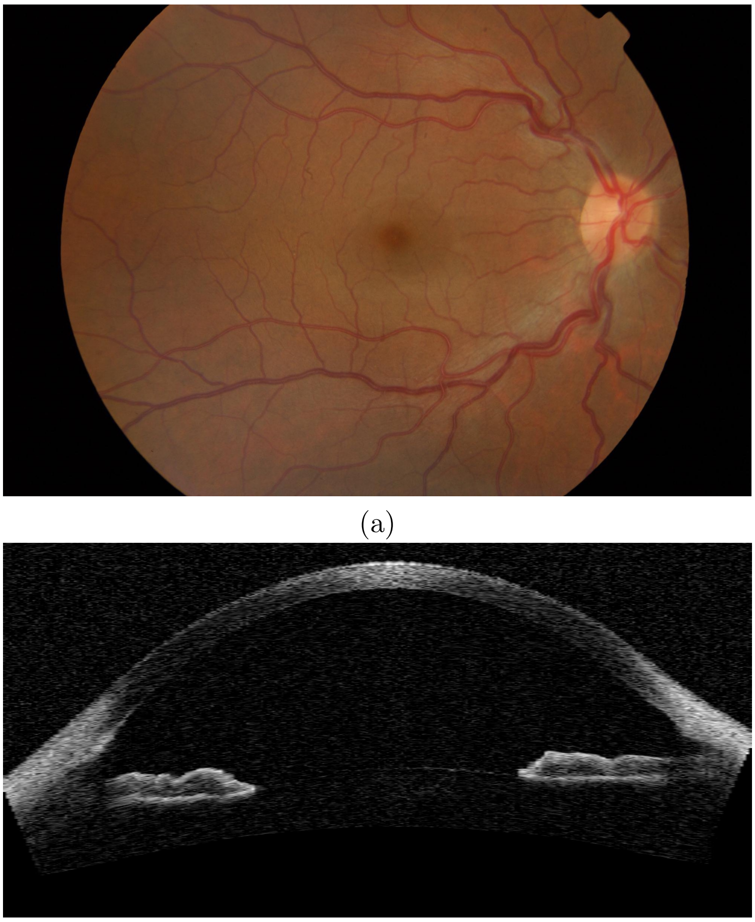
is a simple proxy for understanding the structure of a network. Scale-free networks, for example, are defined as the networks whose degree distribution follows a power law asymptotically.

Weighted networks, are those whose links have a certain weight or strength. Such strength is usually represented with a weight matrix, $\mathbf{w}$, whose entries $\left(w_{i,j}\right)$ represent the weight of the link between nodes i and j.

The concept of weighted degree distribution extends the concept of degree distribution to weighted graphs. In chapter 4 (Ref. [9]) we used it to distinguish between healthy and unhealthy eyes. The weighted degree of a node j in an undirected graph represented by its weight matrix, $\mathbf{w}$, is simply the sum

$$\mathrm{wdeg}\left(j\right)=\sum_{i=1}^{i=N}w_{i,j}.$$

Again, as there are many such weighted degrees as there are nodes, the distribution of the weighted degrees is of interest.



(a)

(b)

Figure 1.4.1.: Examples of ophthalmic images, a color fundus photograph (a) and an anterior chamber OCT (b).

1.4. Datasets analyzed

1.4.1. Ophthalmic images

There is a great variety of techniques for recording ophthalmic images [21–23], including ultrasound, color photograph, optical coherence tomography (OCT) [24, 25], and multispectral images [26]. In this thesis, we have used two different types of ophthalmic images, namely, anterior chamber OCT images, and color fundus photographs. Typical examples are shown in Fig. 1.4.1.

OCT images were studied in chapter 2 (Ref. [7]), and also served as motivation and testing benchmark for chapter 3 (Ref. [8]). These images are recorded by using infra-red light and measuring the time-of-flight delay of the reflected (or backscattered) light to reconstruct the image. This technique can be used either in the posterior chamber of the eye (the retina) or in the anterior chamber of the eye formed by the space between the cornea and the iris. We used the latter in both our studies.

We have used a database of 1213 images taken from patients at IMO (Ocular Microsurgery Institute, Barcelona). The images were acquired using a Visante OCT instrument (Carl Zeiss Meditec) using the "Anterior segment" scan and the "Enhanced anterior segment" scan. The original resolution of all the images is 256-by-1024 pixels corresponding to an area of 16 mm-by-8 mm. All the patients included in the study consented that their data and images could be used for research and teaching purposes and the study was approved by the ethical committee for clinic research in IMO on July 4$^{\text{th}}$, 2017. The patients were selected based on the already available data at IMO (retrospective study). We selected 247 patients at IMO database, 81 of which were glaucoma patients and the rest were a mixed group of healthy patients and patients with other diseases different from glaucoma and with normal intraocular pressure, 104 of those were cataract patients, from which we used images from before and after intraocular lens implant. The average age of the subjects (at the time of the procedure) is 42 years and 57% of such subjects were female.

Color fundus photographs are taken with an optical apparatus that couples the eye with a normal color camera, in order image the retina. This means that, in contrast to OCT imaging which is quite expensive, it can be implemented very cheaply [27]. We used such images in chapter 4 (Ref. [9]) to study the visible vessel network. We used data from three different sources:

1) The High Resolution Fundus database (HRF), a public[1] database [28] which contains 45 color fundus images divided in 15 images of healthy patients, 15 images of patients with diabetic retinopathy and 15 images of glaucomatous patients. These images were captured using a Canon CR-1 fundus camera with a field of view of 60° and have a size of 3504 × 2336 pixels. This database also includes a manual segmentation of the vessel network performed by a human expert.

2) The Messidor Image Database[2], consists of 1200 color fundus images taken with a field of view of 45º and resolutions ranging from 1440 × 960 pixels to 2304 × 1536 pixels [29]. Each image is categorized in one of four groups, corresponding to one of diabetic patients without diabetic retinopathy and three of increasing stages of diabetic retinopathy. We worked with the first 400 images in the database that have good resolution (2240 × 1488 pixels).

3) We also used 93 images from patients at IMO: 23 from healthy subjects and 70 from glaucoma patients taken from the same group as the OCT study. The images have a field of view of 45º and a resolution of 2000 × 1312 pixels. We used images from IMO database from consenting patients. The use of this data was approved by IMO's Ethical committee for clinical study with date October 9th, 2018.

1.4.2. Other datasets

In chapter 3 (Ref. [8]) we developed two algorithms for outlier detection in OCT images that were also used to analyze other datasets: credit card transactions and a face database.

We used a publicly available database of credit card transactions [30–35]. This database contains credit card transactions made in September 2013 by European cardholders. It contains 284807 transactions made in two days, of which 492 correspond to frauds. In order to preserve confidentiality, for each transaction the data set only includes the amount of money in the transaction, a relative time, and 28 features that are the output of a principal component analysis (PCA) of all the other metadata related to the transaction.

In the study of outlier finding techniques (chapter 3 and Ref. [8]) we tested the algorithms in several synthetic data showing different structures. We also used a database[3] with real photographs [36] but added artifacts in a similar manner as in Ref. [37], replacing a square in the image with noise.

[1] Download available at: https://www5.cs.fau.de/research/data/fundus-images/

[2] kindly provided by the Messidor program partners (see http://www.adcis.net/en/DownloadThirdParty/Messidor.html)

[3] http://cam-orl.co.uk/facedatabase.html/

In chapter 5 (Ref. [10]) we simulated the intensity of a semiconductor laser with optical injection [38,39]. The simulated time series, although generated with a 3-dimensional stochastic rate equation model, exhibits highly complex dynamics resembling those of high dimensional complex systems. In certain parameters regions, time series showing extreme events can be obtained. In our study, we used supervised learning techniques to predict the amplitude of the next intensity pulse.

1.5. Image processing

Image processing is a field of computer science that deals with transformations on image data. Usually, digital images are represented in computer memory with arrays of numbers, where each number represents the intensity or luminosity of a given picture element (pixel) and (when dealing with color or multispectral images) a given spectrum. We denote a grayscale (monochromatic) image with a single matrix, $I(i,j)$, whose coordinates $i \in [1,..,N]$ and $j \in [1,..,M]$ correspond to the position in the image. An extra dimension to the matrix is included when dealing with color or multispectral images.

Image processing is the task of converting one image, $I(i,j)$, into a different image,

$$T\left[I\left(i,j\right)\right] = I'\left(i,j\right),$$

with some desired property, such as edge enhancement [40], segmentation [41], blurring, resizing, registration [42] etc. In our work, we used several image processing algorithms. For example in chapter 2 (Ref. [7]) we used them to preprocess the images before applying machine learning algorithms, in chapter 4 (Ref. [9]) we used several algorithms to retrieve the topological information of a vessel network using a color photography.

Image processing methods can be classified into two main types, linear and nonlinear methods.

1.5.1. Linear methods

These methods satisfy the following condition

$$T\left[\alpha I_1\left(i,j\right) + \beta I_2\left(i,j\right)\right] = \alpha T\left[I_1\left(i,j\right)\right] + \beta T\left[I_2\left(i,j\right)\right]. \tag{1.5.1}$$

An example of linear methods are the convolution methods. Convolution methods consists of generating a new image by computing the convolution product of the original image and the convolution kernel. Let $w(k,l)$ with $k \in [-n,..,n]$, $l \in [-m,..,m]$ be the convolution kernel, then the convolution product with an image, $I(i,j)$ is another image defined as

$$I'\left(i,j\right) = \left(w \otimes I\right)\left(i,j\right) = \sum_{k=-n}^{k=n}\sum_{l=-m}^{l=m} w\left(k,l\right) I\left(i+k,j+l\right).$$

It should be noted that when the matrix I is evaluated outside of its range, the zero-padding scheme is usually used, meaning that wherever I is evaluated outside of its range, the value is replaced with a zero.

An simple example of smoothing/blurring kernel is [43]

$$\mathbf{w} = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

This kernel implements a moving average filter, by replacing each pixel intensity with the mean intensity of the 3-by-3 square centered at said pixel.

In chapter 4 (Ref. [9]), we used several convolutional methods to implement a line finding filter.

Another example of linear methods, are frequency (Fourier) domain filters. Such filters are applied in the frequency domain using the (linear) 2d-Fourier transformation defined as

$$\mathcal{F}\left(I\right)\left(k,l\right) = \sum_{i=1}^{i=N}\sum_{j=1}^{j=M} I\left(i,j\right)\exp\left[-2\pi j\left(\frac{ik}{N}+\frac{jl}{M}\right)\right]$$

and then converted back to the spatial domain using the inverse 2d-Fourier transformation

$$\mathcal{F}^{-1}\left(I\right)\left(k,l\right) = \frac{1}{NM}\sum_{i=1}^{i=N}\sum_{j=1}^{j=M} I\left(i,j\right)\exp\left[2\pi j\left(\frac{ik}{N}+\frac{jl}{M}\right)\right].$$

Linear methods can also perform geometrical manipulations to the image, such as translation, rotation and resizing. The simplest geometrical manipulation is the translation parallel to an axis, for example, to move the image p pixels to the right, a new image

$$I'\left(i,j\right) = \begin{cases} 0 & if \quad j \leq p \\ I\left(i,j-p\right) & if \quad j > p \end{cases}$$

can be computed. In chapter 2 (Ref. [7]) we used geometrical transformations to align all the images before analyzing them.

1.5.2. Nonlinear methods

Nonlinear methods are those which do not satisfy Eq. 1.5.1. Segmentation methods are all examples of nonlinear methods, as they output a binary (or multiclass) image.

The simplest example of segmentation method is thresholding, in this case a new image I' is computed using the original image I by applying the rule

$$I'\left(i,j\right) = \begin{cases} 1 & if \quad I\left(i,j\right) \geq t \\ 0 & if \quad I\left(i,j\right) < t \end{cases}$$

where t is the threshold.

In chapter 4 (Ref. [9]) we used a more complex segmentation algorithm proposed in Ref. [41] which agglomerates neighboring pixels if the color difference between them is below a certain threshold, generating pixel clusters that are identified as background or foreground.

Other useful nonlinear methods are edge-preserving smoothing filters that tackle the problem of noise filtering (smoothing) while preserving the relevant edges of the image. A simple example of such filters is the median filter. The n-by-m median filter can be computed by replacing each image value ($I\left(i,j\right)$) with the median of all the values in an n-by-m rectangle centered at the (i,j) point. We used this method in chapter 2 (Ref. [7]) as a first filter to eliminate part of the noise contained in OCT images.

Another useful edge-preserving smoothing filter which we also used in chapter 2 (Ref. [7]) is the anisotropic diffusion [44, 45]. The simulation of an isotropic diffusion process can be used as a linear smoothing filter, if the grayscale image $I(i,j)$ is viewed as the initial concentration of a certain solute, and the processed image, $I'(i,j)$ the concentration after a certain period of time. As the diffusion equation is linear, so is the transformation, resulting in a smoothing filter that doesn't preserve any edges in the original image. In anisitropic diffusion, the diffusion equation is modified by varying the diffusion constant according the gradient of the concentration. Resulting in a nonlinear diffusion equation that produces an edge preserving smoothing filter.

1.6. Thesis outline

This thesis is organized as follows: chapter 2 presents an unsupervised machine learning algorithm to order OCT images in a plane according to the patients risk of developing angle-closure glaucoma. Chapter 3 presents two algorithms, based on graph analysis, to find outliers in multidimensional data. Chapter 4 describes a method to analyze fundus images, that allows to retrieve informative features for differentiating healthy subjects from patients with glaucoma or diabetic retinopathy. Chapter 5 presents the comparison of several machine learning algorithms for forecasting chaotic time series with and without extreme events. In chapter 6 we discuss the results obtained in the previous chapters. Finally, in chapter 7 we draw our conclusions and consider possibilities for future work expanding the results presented in this thesis.

Chapter 2.

Unsupervised feature extraction of anterior chamber OCT images for ordering and classification

In this chapter we propose an image processing method for ordering anterior chamber optical coherence tomography (OCT) images in a fully unsupervised manner. The method consists of three steps: Firstly we preprocess the images (filtering the noise, aligning and normalizing the resolution); secondly, a distance measure between images is computed for every pair of images; thirdly we apply a machine learning algorithm that exploits the distance measure to order the images in a two-dimensional plane. The method is applied to a large ($\sim$1000) database of anterior chamber OCT images of healthy subjects and patients with angle-closure and the resulting unsupervised ordering and classification is validated by two ophthalmologists. The results presented in this chapter are published in Ref. [7].

2.1. Introduction

Machine learning methods are extremely useful in biomedicine [2, 46] and in particular for glaucoma detection [47–50]. The use of such methods can help to optimize the available human resources, to increase accuracy in diagnosis, and to make treatment decisions faster.

Glaucoma is the leading cause of global irreversible blindness [51]. Early diagnosis and treatment is a challenge given that glaucoma presents no symptoms in its early stages [52]. Diagnostic of angle-closure is based on the clinical observation of the angle at the slit-lamp requiring a goniolens that is placed on the patient's cornea. Anterior Segment Optical Coherence Tomography (AS-OCT) is a fast, useful and contact-less tool that allows visualization and measurement of the anterior chamber angle [53–55]. Various techniques exist and are being developed to improve image quality [56–59], and work is also focused on the development of advanced tools to analyze such images [60–62]. However the quality of the images is not always enough for an accurate diagnosis and for example, [63] two glaucoma experts working together were unable to locate the scleral spur (see Fig. 2.4.2) in 28% of the images. Such landmark is of utmost importance to determine most of the relevant features that are used in angle-closure diagnosis.

Previous approaches for processing anterior chamber OCT images used segmentation algorithms and further analysis, while others used manual landmark determination. In works by Tian et al. [60] and Wu et al. [64] segmentation algorithms for OCT anterior chamber images were proposed, both discuss the difficulties of such task due to the noise in the image and to other artifacts.

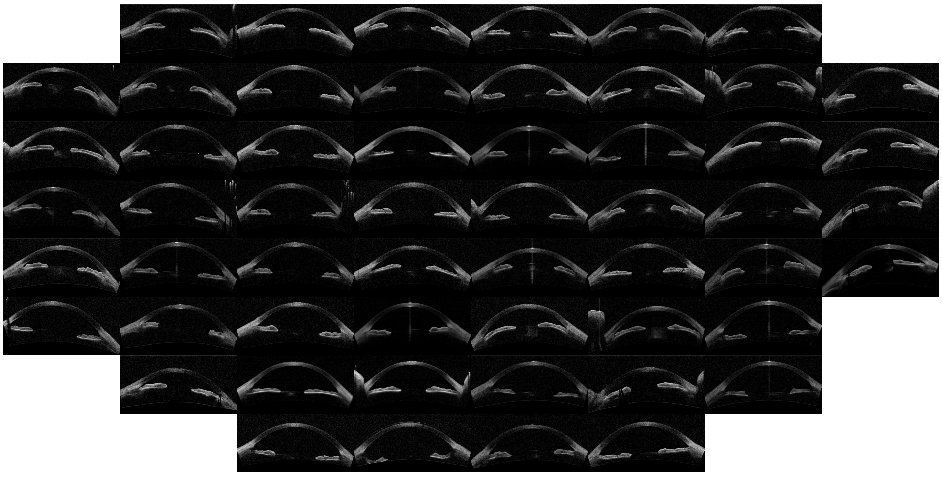


Figure 2.2.1.: Image map obtained when using *IsoMap* and the Euclidean distance, without performing the alignment step in the pre-processing of the images.

In Ref. [65] the authors analyzed data from manual landmark determination to classify images into five subcategories of angle-closure glaucoma. They used both supervised and unsupervised feature selection and AdaBoost classifiers (supervised learning) to achieve accuracies in the range of 84%∼87%. In Ref. [66] a method was proposed to extract almost 3000 features from the raw images, then, the most relevant features for classification were supervisedly selected, and machine learning was applied to the selected features. However, the fact that only 74 images were analyzed while 15 features were used, compromised the statistical significance of the study. In Ref. [67] Histogram of Oriented Gradients (HOG) features of anterior chamber OCT images and Support Vector Machine (SVM, supervised learning) were used to classify different glaucoma subtypes, achieving accuracies in the order of 80%.

In this chapter we propose an image processing method that *unsupervisedly* orders OCT anterior chamber images according to what we demonstrate to be relevant extracted features. The reliability of the algorithm is tested with a large number of images (∼1000). Importantly, our method is fully autonomous and can be used to analyze images with a wide spectrum of quality, even those with high levels of noise and artifacts.

2.2. Results

The outcome of the algorithm applied to the image database (both described in Sec. 2.4) is presented in the way of an "Image Map". A regular grid in the coordinates space (w, v) is defined and one image per grid point is displayed. The results are presented in Figs. 2.2.1-2.2.4 that show the Image Map obtained after applying *IsoMap* to the Euclidean distance (Fig. 2.2.1), to the aligned Euclidean distance (Eq. 2.4.4, Fig. 2.2.2), to the aligned Hellinger distance (Eq. 2.4.3, Fig. 2.2.3) and after applying *t-SNE* to the Hellinger distance (Fig. 2.2.4). These figures are ordered with increasing complexity and performance of the corresponding algorithm.

In Fig. 2.2.1 (*IsoMap* with Euclidean distance) it is apparent that the algorithm ordered the images according to the orientation (horizontal axis) and position (vertical axis) of each eye inside the OCT image, which are irrelevant features, this simple algorithm is, then, not capable of extracting useful features. However, when including the alignment step in the pre-processing of the images

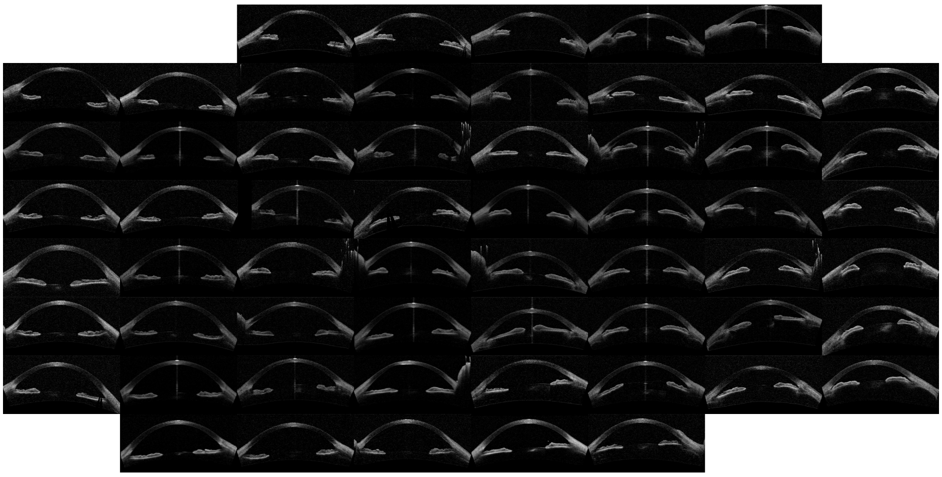


Figure 2.2.2.: Image map with results using *IsoMap* and the Euclidean distance, including the alignment step in the pre-processing of the images.

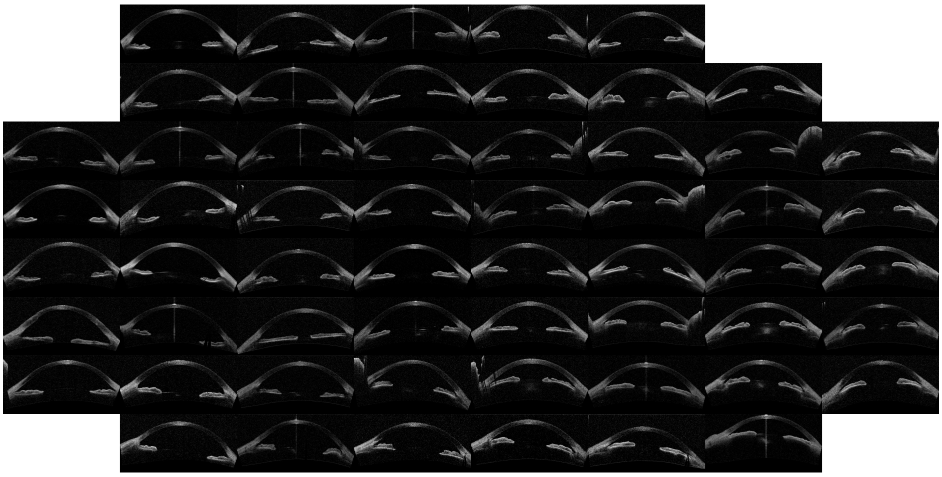


Figure 2.2.3.: Image map with results using *IsoMap* and the Hellinger distance, including the alignment step in the pre-processing of the images.

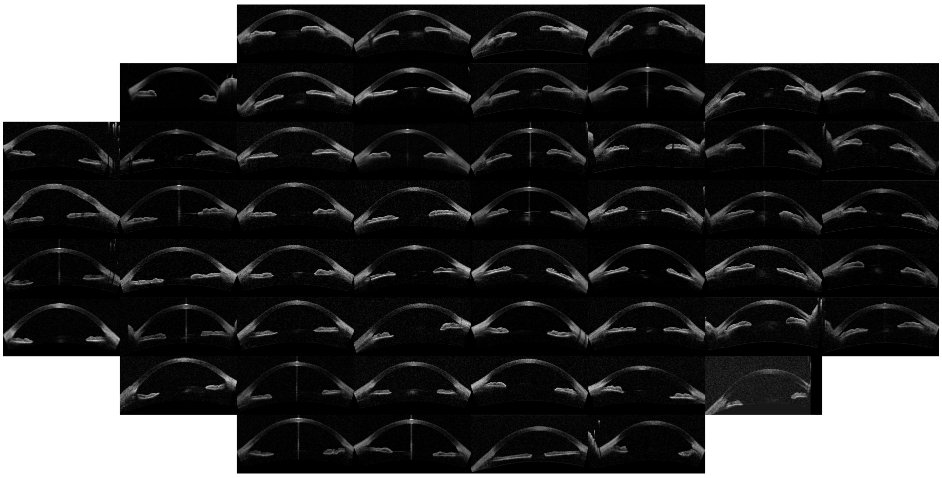


Figure 2.2.4.: Image map with results using *t-SNE* and the Hellinger distance, including the alignment step in the pre-processing of the images.

which removed the variability detected by the first algorithm, meaningful features were extracted, as it can be seen in the Image Map in Fig. 2.2.2 (these features correlate with the features derived manually, as shown in Table 2.1). A similar map was obtained with the Hellinger distance, shown in Fig. 2.2.3 which marginally improves the performance with the Euclidean distance. To test the robustness of the image ordering, the *t-SNE* algorithm was applied (instead of *IsoMap*). The map obtained is shown in Fig. 2.2.4, which turned out to perform slightly better than *IsoMap* according to the correlations shown in Table 2.1.

In Fig. 2.2.5 the features returned by the unsupervised algorithm, using *t-SNE*, are compared with the features obtained from the manual annotation of the images: in panel (a) the color code indicates the chamber depth, and in panel (b), it indicates the mean angle (average of α and β) of each annotated image. Clearly, the features obtained from the manual annotation correlate very well with the features returned by the unsupervised algorithm. However, one should notice that the annotated features are not independent, but strongly correlated with each other. A summary of the correlation between the ordering (Mappings) and different features is shown in Table 2.1.

Method	Mean angle	Chamber Depth	Manual classification
IsoMap, no alignment	0.05	0.07	0.18
IsoMap, aligned euclidean	0.79	0.88	0.77
IsoMap, Hellinger	0.80	0.89	0.79
t-SNE, Hellinger	0.80	0.90	0.81

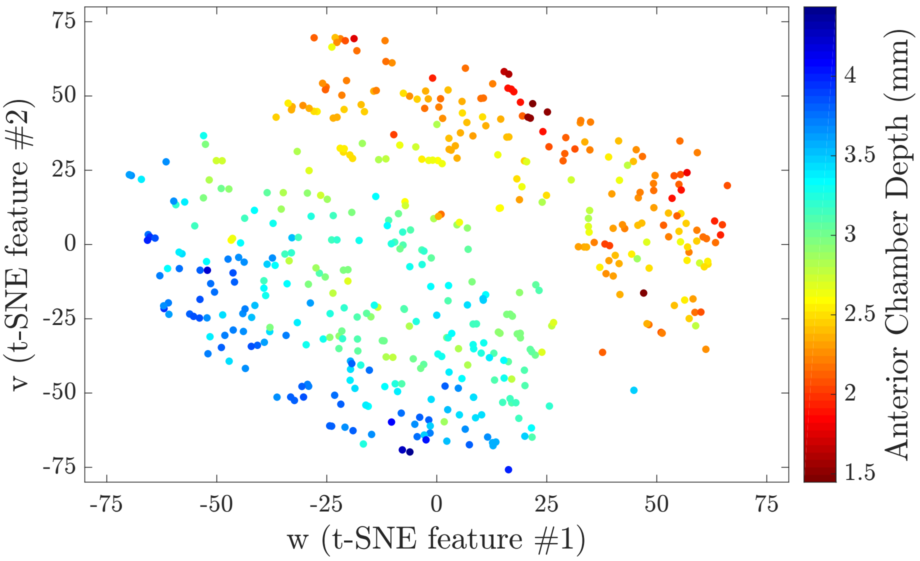
Table 2.1.: Correlation coefficient between the automatic feature extracted (selecting the direction in the mapped space that gives the larger correlation) and the manual annotation features (mean angle and depth). We also present the correlation with manual labels (0 corresponds to wide open, 1 to open, 2 to narrow, and 3 to closed). For comparison the correlation coefficients of the manual labels and anterior chamber depth is 0.8 while with ARA_500 (see Ref. [65]) is 0.75.

Finally, the manual classification done by two expert ophthalmologists is included in the *t-SNE* (w, v) map. In Fig. 2.2.6.(a) the color code indicates the different classes, and it can be noticed that the data points that represent images in the "wide-open" category are scattered in the left-down corner of the map, while the data points that represent images in the "closed" category are scattered in the right-top corner of the map.

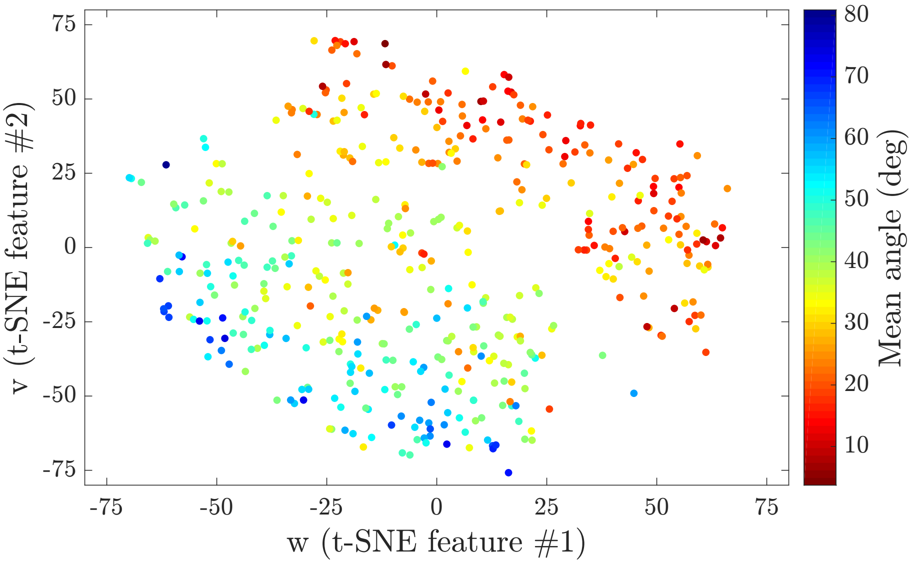
In order to demonstrate that the separation between different classes obtained from features retrieved from the manual annotation of the images is comparable to the one obtained from the *t-SNE* features, the manual annotation features (mean angle and smallest ARA_500) are plotted in Fig. 2.2.6.(b). A comparison with panel (a) reveals that the manual annotation features also do not allow for a clear-cut separation of the different classes, and therefore, for classification purposes, the features obtained with the *t-SNE* map are as good as the manual annotation features.

2.3. Discussion

We have proposed a new algorithm for ordering anterior chamber OCT images in such a way that it is possible to classify them, in a fully unsupervised manner, in meaningful groups according to relevant features. We have tested the algorithm with a large set of images classified by two expert ophthalmologists, and with a larger set of annotated images. We have verified that the separation

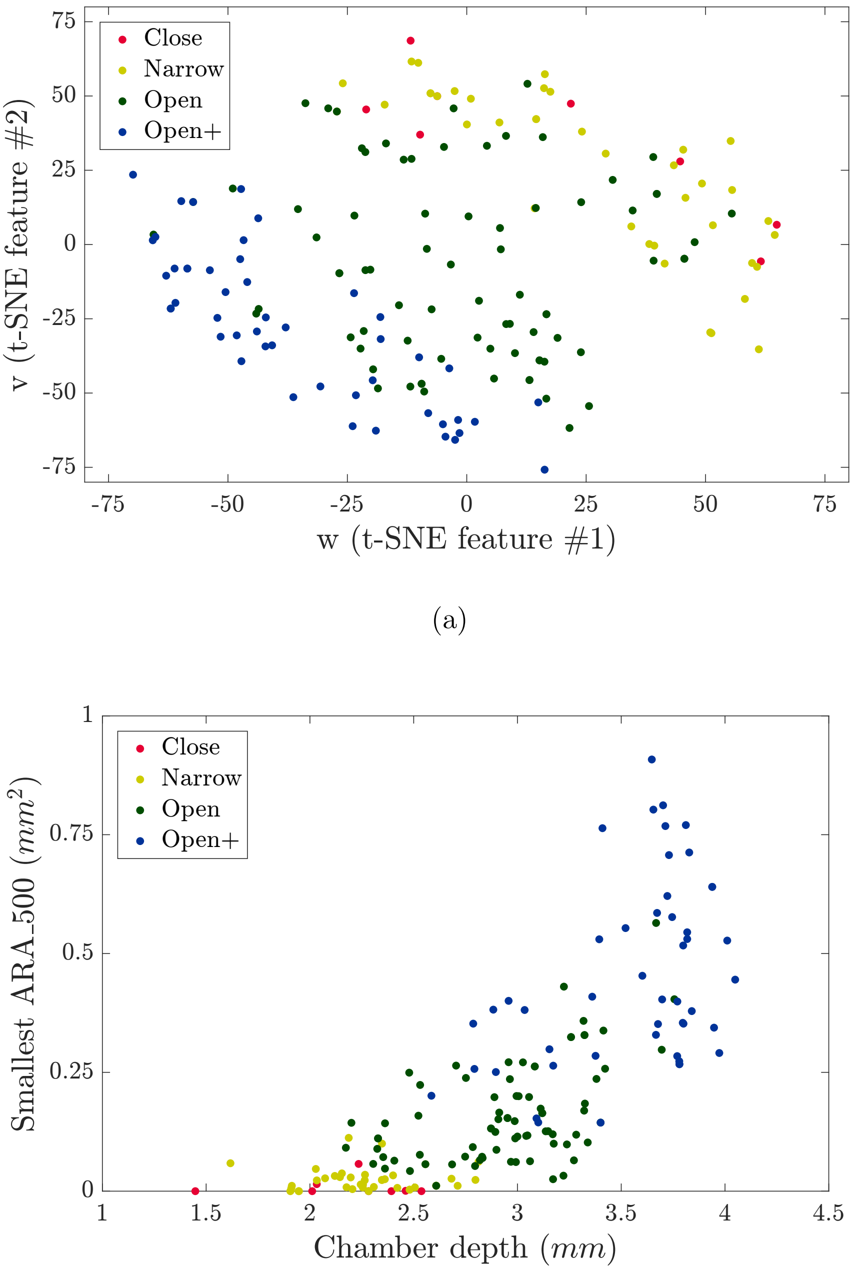


(a)



(b)

Figure 2.2.5.: (a) Comparison between t-SNE results and chamber depth from manually annotated images. (b) Comparison between t-SNE results and mean angle from manually annotated images.



(a)

(b)

Figure 2.2.6.: (a) Comparison between t-SNE results and manual classification. (b) Comparison between annotated results (angle and smallest ARA_500, Angle Recess Area see Ref. [65]) and manual classification.

in the different classes defined by the ophthalmologists (closed, narrow, open, and wide open) is similar when using the manually extracted features, or when using the features that are returned by the unsupervised algorithm (Fig. 2.2.6).

Therefore, the abstract features generated by the algorithm provide novel tools for assessing OCT images of the anterior chamber. They can be used for direct classification of the images and, furthermore, they can be linked to established quantities used for characterizing diseased eyes (like chamber depth, iris-corneal angle) resulting in an automatic detection system. As the algorithm is fully unsupervised, it can be easily automated and set up in OCT imaging systems to aid technicians and doctors in an early diagnosis.

The two main advantages of the algorithm demonstrated here over previous works are that it doesn't need any ground truth or gold standard for training, and it does not rely on specific landmarks; thus, it can analyze images in which relevant landmarks are not visible or not easy to locate.

2.4. Methods

2.4.1. Data

The data consists of 1213 OCT images taken from consenting patients at IMO (Ocular Microsurgery Institute, Barcelona [68]). The images were acquired using a Visante OCT instrument (Carl Zeiss Meditec) using the "Anterior segment" scan and the "Enhanced anterior segment" scan. The original resolution of all the images is 256-by-1024 pixels corresponding to an area of 16 mm-by-8 mm. All the patients included in the study consented that their data and images could be used for research and teaching purposes and the study has been approved by the ethical committee for clinic research in IMO. The patients were selected based on the already available data at IMO (retrospective study). We selected 247 patients at IMO database, 81 of which were glaucoma patents and the rest was a mixed group of healthy patients and patients with other diseases different from glaucoma and with normal intraocular pressure, 104 of those were cataract patients, from which we used images from before and after intraocular lens implant (images with IOL implants were treated in the same manner as the rest of the images). The average age of the subjects (at the time of the procedure) is 42 years and 57% of such subjects are female. We retrieved all the available images of the selected patients that corresponded to the mentioned scans. We had to discard around 1400 images because they didn't depict the whole anterior segment (they were deliberately zoomed in) and around 50 due to very poor quality.

2.4.2. Manual annotation and classification of images

Two glaucoma experts (Elena Arrondo, MD and Cecilia Salinas, MD) evaluated a subset of 160 images and classified them into four categories: closed, narrow, open, and wide open. An image example in each category is displayed in Fig. 2.4.1. It is important to remark that this manual classification was not used by the algorithm (as it does not "learn" and thus, it does not require any training set); the manual classification was only used to test the relevance of the features returned by the nonlinear dimensionality reduction algorithms.

To test the relevance of the features extracted by the algorithm, several relevant landmarks were manually annotated in a subset (∼400) of the images. The landmarks used are (see Fig. 2.4.2):

- Scleral spur.
- A second point near the scleral spur in the inside edge of the cornea (to set a line approximating the inside edge of the cornea).
- Two points on the top edge of the iris.
- Points in the inside and outside edge of the cornea in the center
- A point in the top edge of the lens in the center.

From those landmarks the following features were calculated:

- Anterior chamber depth (L).
- Iris-corneal angles (α and β).
- Angle recess area (ARA_500, see Ref. [65])

It has to be noted that some landmarks were not always clearly visible, in such cases the human expert guessed its position based on nearby features. The most common landmark that had to be guessed was the point in the top edge of the lens, which can be guessed based on the position of the iris. The sclerar spur is also frequently difficult to find but usually the clues were enough to guess its position by the expert. Ultimately, if the expert thought that it wasn't possible to make a good guess, the image was simply skipped and omitted from the subset of manually annotated images. As with the classification, these landmarks were not used by the algorithm, rather they were used afterwards to evaluate its results.

2.4.3. Unsupervised ordering and classification algorithm

In this section we present the algorithm for the unsupervised ordering and classification of OCT images. The input to the algorithm is a database of anterior chamber OCT images and the output is a map in a two dimensional plane. The algorithm performs three main steps: pre-processes the images, calculates a pair-wise distance measure between images and applies nonlinear dimensionality reduction.

2.4.3.1. Image pre-processing

The pre-processing of the images consists of the following substeps: homogenization, filtering, centering and aligning.

Homogenization: in each image, the intensity of each pixel was converted to double precision and normalized to be in between 0 and 1 (by linearly rescaling). Then, the horizontal and vertical spatial resolutions were adjusted to be the same (note that the original spacial resolution is anisotropic).

Filtering: First, a two dimensional rectangular median filter was applied to the image (with a 0.055mm-by-0.117mm rectangle). This filter was needed because of the process of adjusting the spatial resolution, which resulted in the noise being spread more in one direction than in the other. Then, an anisotropic diffusion [44, 45] filter was applied to smooth the image, removing the spatial high-frequencies while preserving relevant edges. An example of such filtering is shown in Fig. 2.4.3.

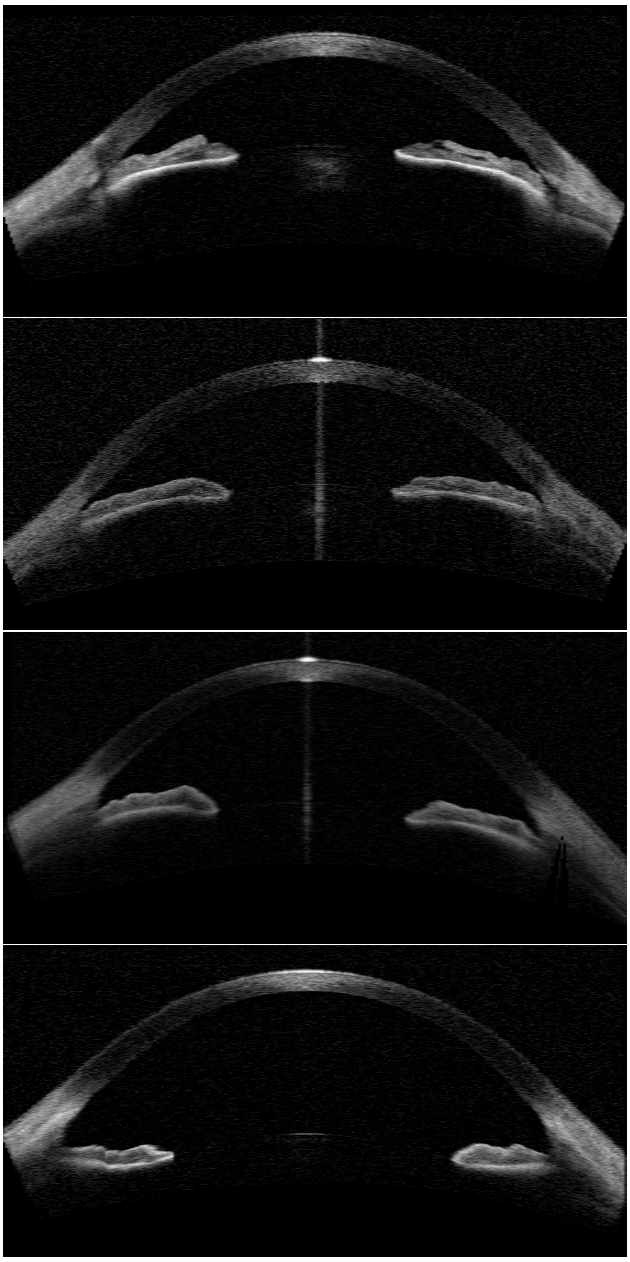


Figure 2.4.1.: Example images of the categories classified by glaucoma experts. Closed (top), narrow (middle-top), open (middle-bottom), and wide open (bottom).

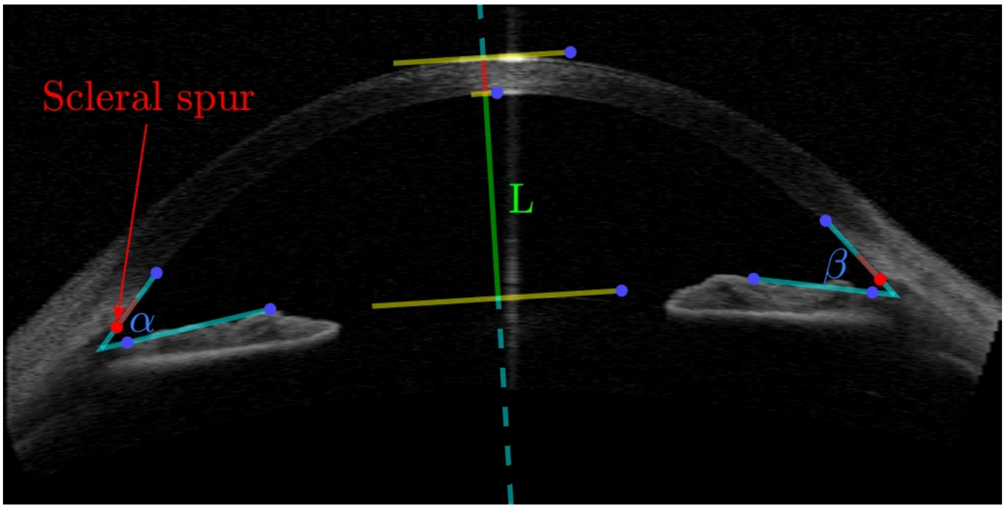


Figure 2.4.2.: Example of an annotated image with landmarks and guidelines.

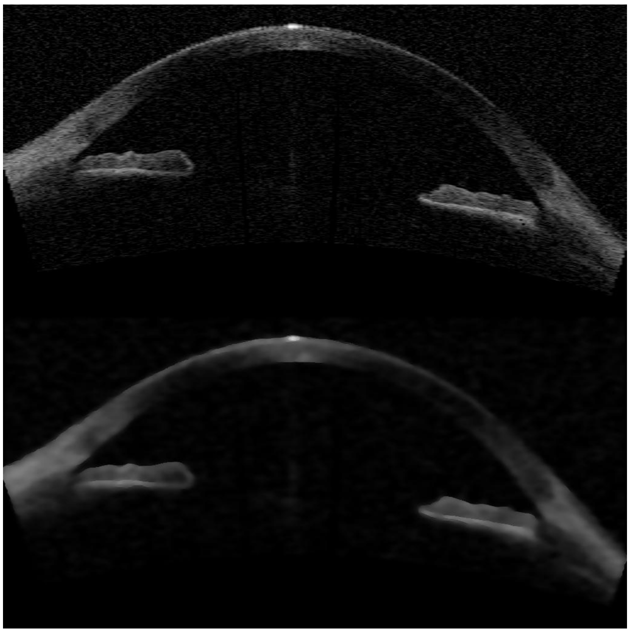


Figure 2.4.3.: Example raw OCT image (top), and the same image after the filtering process (bottom).

For centering and aligning a set of statistical properties of the images were calculated, namely:

$$\begin{array}{rcl} S=\sum_{i,j}M\left(i,j\right) & , & X=\sum_{i,j}jM\left(i,j\right)\ , \\ Y=\sum_{i,j}iM\left(i,j\right) & , & XX=\sum_{i,j}j^{2}M\left(i,j\right)\ , \\ YY=\sum_{i,j}i^{2}M\left(i,j\right) & , & XY=\sum_{i,j}ijM\left(i,j\right)\ , \end{array} \tag{2.4.1}$$

where $M\left(i,j\right)$ is the (gray) value of the image on the pixel that is i pixels down from the top edge and j pixels right from the left edge. With these quantities the centroid of the image (whose coordinates are $(i,j)=\left(\frac{Y}{S},\frac{X}{S}\right)$) and the covariance matrix (COV) were calculated:

$$\text{COV}=\begin{pmatrix} \frac{YY}{S}-\frac{Y^2}{S^2} & \frac{XY}{S}-\left(\frac{X}{S}\right)\left(\frac{Y}{S}\right) \\ \frac{XY}{S}-\left(\frac{X}{S}\right)\left(\frac{Y}{S}\right) & \frac{XX}{S}-\frac{X^2}{S^2} \end{pmatrix}\ . \tag{2.4.2}$$

From the covariance matrix the eigenvector v_1, corresponding to the largest eigenvalue, was calculated and used to generate a new image, $M^{(C)}$, twice as large as M, that was initialized with zeros. Then, M was copied to $M^{(C)}$ such that the centroid of M coincides with the center of $M^{(C)}$ and v_1 is aligned with the horizontal direction. The elements of $M^{(C)}$ which were not overwritten with the elements of M remained zero.

2.4.3.2. Pair-wise distance measure between images

In this step a distance matrix $(D\left(l,m\right))$ was calculated whose entries are the pair-wise distances between images l and m. Two distance measures were employed: the Hellinger distance (d_H) [69],

$$D\left(l,m\right)=d_H\left(M_l^{(C)},M_m^{(C)}\right)=\sqrt{2\sum_{i,j}\left(\sqrt{\frac{M_l^{(C)}\left(i,j\right)}{S_l}}-\sqrt{\frac{M_m^{(C)}\left(i,j\right)}{S_m}}\right)^2}\ . \tag{2.4.3}$$

and the Euclidean distance,

$$D\left(l,m\right)=d_E\left(M_l^{(C)},M_m^{(C)}\right)=\sqrt{\sum_{i,j}\left(M_l^{(C)}\left(i,j\right)-M_m^{(C)}\left(i,j\right)\right)^2}\ . \tag{2.4.4}$$

2.4.3.3. Nonlinear dimensionality reduction

In order to extract meaningful information from the pair-wise distance matrix, $D\left(l,m\right)$, we applied a nonlinear dimensionality reduction algorithm, directly to the pair-wise distance matrix. Two algorithms were tested: *IsoMap* [11] and *t-SNE* [15].

These algorithms assign, to each image in the database, a point in a n-dimensional space, whose coordinates will be referred to as mapped coordinates. While the algorithm also works with an arbitrary number of dimensions, in this chapter a two dimensional space (w,v) is used for visualization reasons. The choice of a two dimensional space is appropriated because the residual variance (defined as in Ref. [11]) using two dimensions is of about 30%.

2.4.3.4. Computational runtime

All the described algorithms were implemented and run using MatLab in a portable computer with an Intel i7-7700HQ processor and 16GB of RAM. We used the implementation of the nonlinear dimensionality reduction techniques written by van der Maaten et al. [70]. With this setup, it takes 5294 seconds (one and a half hours) to preprocess all the (1213) images including aligning and filtering, 1054 seconds (18 minutes) to compute the Hellinger distance matrix (735078 pair-wise distances), it takes 41 seconds for IsoMap to compute the mapping, and it takes 25 seconds for t-SNE to compute the mapping. It must be noted that all this runtimes could be significantly improved by rewriting the algorithms in a compiled language.

Informed consent statement: informed consent statements were obtained from all the participants of the study.

Guidelines and regulations: OCT image acquisition and analysis were performed in accordance with the relevant European guidelines and regulations.

Chapter 3.

Outlier mining methods based on graph structure analysis

Outlier detection in high-dimensional datasets is a fundamental and challenging problem across disciplines that has also practical implications, as removing outliers from the training set improves the performance of machine learning algorithms. While many outlier mining algorithms have been proposed in the literature, they tend to be valid or efficient for specific types of datasets (time series, images, videos, etc.). Here we propose two methods that can be applied to generic datasets, as long as there is a meaningful measure of distance between pairs of elements of the dataset. Both methods start by defining a graph, where the nodes are the elements of the dataset, and the links have associated weights that are the distances between the nodes. Then, the first method assigns an outlier score based on the percolation (i.e., the fragmentation) of the graph. The second method uses the popular IsoMap nonlinear dimensionality reduction algorithm, and assigns an outlier score by comparing the geodesic distances with the distances in the reduced space. We test these algorithms on real and synthetic datasets and show that they either outperform, or perform on par with other popular outlier detection methods. A main advantage of the percolation method is that is parameter free and therefore, it does not require any training; on the other hand, the IsoMap method has two integer number parameters, and when they are appropriately selected, the method performs similar to or better than all the other methods tested. The results presented in this chapter are published in Ref. [8].

3.1. Introduction

When working with large databases, it is common to have entries that may not belong to the database. Sometimes this is because they were mislabeled, or some automatic process failed and introduced artifacts. On the other hand, anomalous items that appear not to belong, may actually be legitimate, just extreme cases of the variability of a large sample. All these elements are usually referred to as outliers [71, 72]. In general, outliers are observations that appear to have been generated by a different process than that of the other (normal) observations.

There are many definitions of what an outlier is, which vary with the system under consideration. For example, rogue waves (or freak waves), which are extremely high waves that might have different generating mechanisms than normal waves [73], have been studied in many fields [74–78], including hydrodynamics and optics. They are usually defined as the extremes in the tail of the distribution of wave heights, however, their precise definition varies, as in hydrodynamics a wave whose height is larger than three times the average can be considered extreme, while in optics, much higher waves compared to the average can be observed [79].

In the field of computer science, a practical definition of outlier elements is that they are those elements that, when they are removed from the training data set, the performance of a machine learning algorithm improves [80]. Outlier mining allows to identify and eliminate mislabeled data [81, 82]. In other situations, the outliers are the interesting points, for example to perform fraud detection [83, 84] or novelty detection [85]. The terms novelty detection, outlier detection and anomaly detection are sometimes used as synonyms in the literature [85, 86].

In spatial objects, the identification of anomalous regions that have distinct features from those of their surrounding regions can reveal valuable information [87–89]. This is the case of biomedical images where particular anomalies characterize the presence of a disease [90, 91]. For example, Schlegl et al. [92] recently proposed a generative adversarial network for detecting anomalies in OCT retinal images. Another relevant problem consists in anomaly detection in sequences of ordered events, a comprehensive review was provided in Chandola et al. [93], where three main types of formulations of the problem were identified: i) to determine if a given sequence is anomalous with respect to a database of sequences; ii) to determine if a particular segment is anomalous within a sequence and iii) to determine if the frequency of given event of sequence of events is anomalous with respect to the expected frequency.

With increasing computer power, neural networks are also an attractive option for detecting outliers [94, 95] and anomalies [96]. Hodge and Austin [72] have classified outlier detection methods in three groups: unsupervised (methods that use no prior knowledge of the data), supervised (methods which model both normal and outlier points), and semi-supervised (methods that model only normal points, or only outliers), although the latter can also include a broader spectrum of algorithms (for example a combination of fully unsupervised method and a supervised one). A recent review of outlier definitions and detection methods is presented in Zimek and Filzmoser [97].

We are interested in outlier detection in data that belong to a metric space [98–101]. In this type of dataset, a distance can be defined between items. A relevant example is a wireless sensor network, where localization is based on the distances between nodes and the presence of outliers in data results in localization inaccuracy [102, 103]. Abukhalaf et al. [104] presents a comprehensive survey of outlier detection techniques for localization in wireless sensor networks.

Here we propose two methods that use, as input, only the distances between items in the dataset. Both methods define a graph, or a network, where the nodes are the items of the dataset, and the links have associated weights which are the distances. Then, each method identifies outliers by analyzing the structure of the graph. The first method assigns to each item an outlier score based on the percolation (i.e., the fragmentation) of the graph. The second method uses the IsoMap algorithm [11] (a nonlinear dimensionality reduction algorithm that learns the manifold in which the data is embedded in a reduced space), and assigns to each element an outlier score by comparing the geodesic distances with the distances in the reduced space.

Numerous algorithms have been proposed in the literature that use manifold embedding, or more in general, graph embedding, either explicitly or implicitly, to detect anomalies in data [105–110]. A comprehensive review of the literature is out of the scope of the present work, but here we discuss a few relevant examples. Agovic et al. [111] and Wang et al. [112] used the IsoMap algorithm as a preprocessing step, before applying the actual outlier finding algorithm. Our approach differs fundamentally because we take into account how well or how poorly items fit in the manifold, which is disregarded by the cited methods, as they only perform outlier detection in the reduced space.

In Brito et al. [113] the authors use the distance matrix to build a graph where two nodes are connected if each of them is between the k's closest neighbors. For a sufficiently large value of k, the graph will be connected, while, for small values of k, disjoint clusters will appear. If the clusters that appear are large enough, they are considered as classes, while if they are small, they can be interpreted as outliers. In contrast to traditional k-NN algorithms, where the number of neighbors has to be determined a priori, the method proposed by Brito et al. [113] finds the value of k automatically. Nevertheless, the method is not truly parameter-free, as there are two parameters that have to be adjusted which depend on both the dimension and size of the dataset. We speculate that this graph fragmentation method identifies similar outliers as our percolation method, which has the advantage of being parameter free.

We demonstrate the validity of the percolation and IsoMap methods using several datasets, among them, a database of optical coherence tomography (OCT) images of the anterior chamber of the eye. OCT anterior chamber images are routinely used for the early diagnosis of glaucoma. We show that, when images with artifacts (outliers) are removed from the training dataset, the performance of the unsupervised ordering algorithm [7] improves significantly. We also compare the performance of these methods with the performance of other popular methods used in the literature. We show that our results are at worst comparable to those methods.

The chapter is organized as follows, in section 3.2 we describe the proposed methods and also, other popular methods that we use for comparison. In section 3.3, we describe the datasets analyzed. In section 3.4 we present the results and in section 3.5, we summarize our conclusions.

3.2. Methods

In this section we describe the two proposed methods, which we refer to as percolation-based method and IsoMap-based method. Both methods require the definition of a distance measure between pairs of elements of the dataset. We also describe three other outlier mining methods, which we used for comparison.

We consider a dataset with N elements and let i and j be two elements, which have associated vectors with m features, $V_i = \{v_1^i \dots v_m^i\}$ and $V_j = \{v_1^j \dots v_m^j\}$. The distance between these elements can be defined as

$$D_{ij} = \left(\sum_k \left| v_k^i - v_k^j \right|^p \right)^{1/p} \tag{3.2.1}$$

with p an integer number, taken equal to 2 (Euclidian distance) unless otherwise stated. The selection of an appropriate distance measure is of the utmost importance, since it must capture the similarities and differences of the data. Adding a preprocessing step before calculating the distance matrix may also be necessary to obtain significant distances.

3.2.1. Percolation-based method

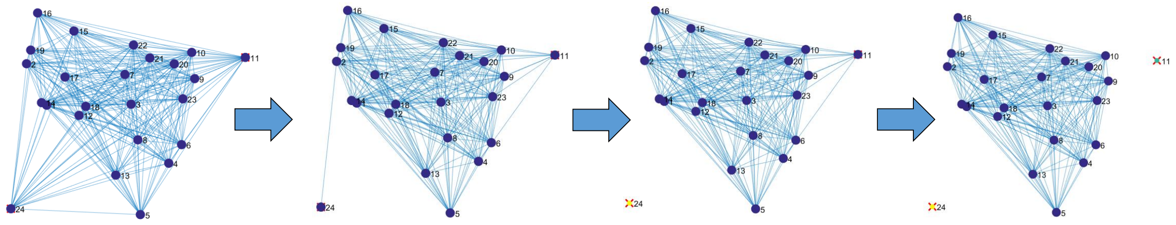
The method is described in Fig. 3.2.1. We begin by considering a fully connected graph, where the nodes are the elements of the set and where the links are weighted by the distance matrix D_{ij}. Now, we proceed in the following way: we remove the links one by one, from higher to lower weights (i.e., the link representing the highest distance between a pair of elements is removed first).

If only a few links are removed, the graph will remain connected, but if one continues, the graph will start to break into different components. As it is well known from percolation theory [114, 115], it is expected for most of the nodes to remain connected inside a single *giant connected component* (GCC), and for the rest of them to distribute into many small components. If we remove enough links, even the giant component disappears. This transition between the existence and non-existence of a giant component is known as a percolation transition, and is one of the most studied problems of statistical physics [116, 117]. Here, we are interested in the percolated state, i.e, when such a giant component exists. In particular, the nodes that do not belong to the GCC are candidates for being considered as outliers, as they are relatively distant to the rest of the graph.

Following this idea, we can label each node with an outlier score (OS), defined as the weight of the link that, after being removed, separates the node from the GCC. Thus, the first elements to leave the GCC are the ones with the highest OS, while the last ones have the lowest OS.

For this method to correctly identify the outliers, we assume that normal points occupy more densely populated zones than outliers, thus having (normal points) local neighborhoods connected with small distances while outliers are connected to normal points via longer distances. Such outliers will become disconnected from the giant connected component sooner than the normal ones in the described procedure.

It is worth noting that the computation of the GCC can be performed efficiently using a variation of the union-find algorithm [118], thus making this method suitable for large datasets.



(a) (b) (c) (d)

Figure 3.2.1.: Example of application of the percolation method. Starting with a fully connected graph (a) links are removed according to their distances (longer distances are removed first). In (b) some of the longest connections have been eliminated, but the graph remains fully connected, in (c) the first outlier is identified as the first element that becomes disconnected from the giant connected component, and in (d) two elements are disconnected.

3.2.2. IsoMap-based method

The basic idea of this method is to use the well-known algorithm IsoMap [11] to perform dimensionality reduction on the raw data, and to analyze the manifold structure in the reduced space, assigning to each point an outlier score that measures how well it fits in the manifold.

The method consists of the following steps

- We apply IsoMap to the distance matrix D_{ij} (computed from the raw features) and obtain two matrices: 1) a new set of features for each element of the database, $V'_i = \left\{ v'^i_1 ... v'^i_r \right\}$ with $i = 1 \ldots N$ and 2) a matrix of graph distances, D^G_{ij} in the geodesic space as described in Tenenbaum et al. [11].

- Using the new set of features, we calculate a new distance matrix $\tilde{D}_{ij}$, using the Euclidean distance (Eq. 3.2.1 with $p = 2$).

- The third step is to compare $\tilde{D}_{ij}$ with D^G_{ij}: for each element i we compute the similarity, ρ_i, between vectors $(D^G_{i1}, \dots D^G_{iN})$ and $(\tilde{D}_{i1}, \dots \tilde{D}_{iN})$, using the Pearson correlation coefficient.

- The final step is to define the outlier score as $OS_i = 1 - {\rho_i}^2$. For "normal" elements, we expect high similarity, while for abnormal ones, we expect low similarity.

With this method, the assumption is that normal points lie in a low dimensional manifold embedded in the full-dimensional space, and outliers lie outside such manifold. If the parameters of the IsoMap are such that the low dimensional manifold structure is recovered successfully, the distances between points in the new set of features ($\tilde{D}_{ij}$), the geodesic distances in the manifold, and the graph distances (D^G_{ij} an approximation of the geodesic distance) should all be similar for normal points lying on the manifold. However, for outliers the geodesic distance is not defined and thus, the graph distances and the distances in the new set of features will disagree. When we compute the similarity, ρ_i, assessing this disagreement, normal points will have a high value ρ_i (near 1) and outliers a low value of ρ_i, therefore the outlier score should be high for outliers and low for normal points.

The parameters of this method, are the parameters of the IsoMap algorithm, namely, the dimensionality of the objective space (d) and neighborhood size (number of neighbors, k) to construct the graph. In this work, the parameters of the IsoMap were optimized (when a training set was available) by maximizing the average precision doing an extensive search in the parameter space.

3.2.3. Other methods

We compared the performance of both methods with:

- The simplest way to define an outlier score: the distance to the center-of-mass (d2CM) in the original feature space, $V_i = \{v^i_1 \dots v^i_m\}$. For "normal" elements, we expect short distance, while for abnormal ones, we expect high distance.

- A popular distance-based method, which will be referred to as Ramaswamy [99]. This method is based on the distance of a point from its kth nearest neighbor, in the raw (original) high-dimensional feature space. The method assigns an outlier score to each point equal to its distance to its kth nearest neighbor.

- And a very popular method, One Class Support Vector Machine (OCSVM) which uses the inner product between the elements in the database to estimate a function that is positive in a subset of the input space where elements are likely to be found, and negative otherwise [119].

3.2.4. Implementation

All the methods were implemented and run in MatLab. The IsoMap method was build modifying the IsoMap algorithm implementation by Van Der Maaten et al. [70], the percolation method was implemented using graph objects in MatLab. With a simple database of 1000 elements with 30 dimensions, the percolation method takes around 6 seconds to run and the IsoMap method takes around 18 seconds, while One Class Support Vector Machine takes around 0.2 seconds to run, Ramaswamy about 0.04 seconds to run, and distance to center of mass 0.01 seconds to run on an Intel i7-7700HQ laptop. Both methods could significantly improve their runtime by optimizing the code and translating it into a compiled language.

3.3. Data

We tested the above described methods in several databases. In this chapter we present three examples: a database of anterior chamber Optical Coherent Tomography (OCT) images, a database of face images with added artifacts, and a database of credit card transactions. Additional synthetic examples are presented in appendix A.

3.3.1. Anterior chamber OCT images

This database consists of 1213 OCT images of the anterior chamber of the eye of healthy and non-healthy patients of the *Instituto de Microcirugia Ocular* in Barcelona. The database was analyzed in chapter 2 (Ref. [7]) where an unsupervised algorithm for ordering the images was proposed. The images had been classified in four categories (closed, narrow, open, and wide open) by two expert ophthalmologists. By using manually extracted features, and the features returned by the unsupervised algorithm, a similar separation in the four classes was found. Here we will demonstrate that the similarity is further improved when images containing artifacts (outliers) are removed from the dataset given to the unsupervised algorithm.

Examples taken from the database are shown in Fig. 3.3.1

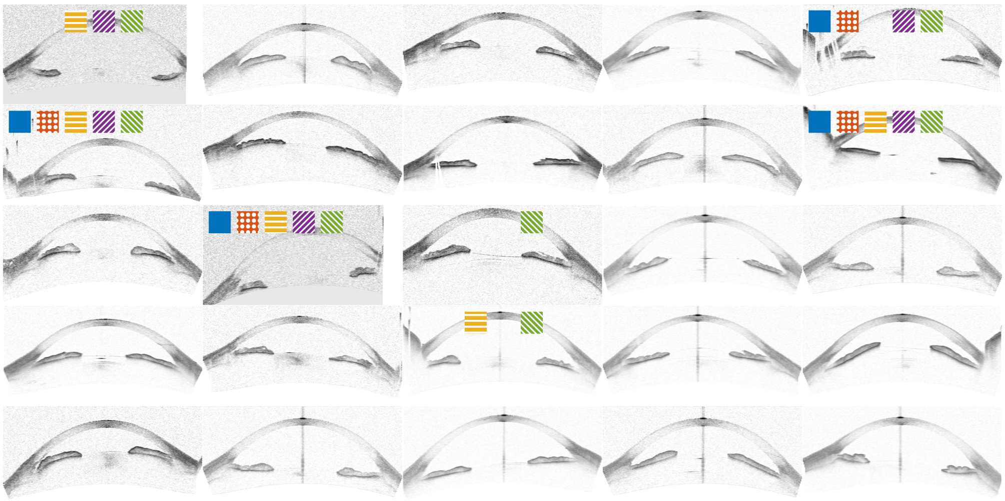


Figure 3.3.1.: Example images from the OCT database, all except the first one were randomly sampled. Marked images correspond to top 15% outlier score for OCSVM (Blue), distance to center of mass (Orange), IsoMap (Yellow), Percolation (Purple), and Ramaswamy (Green). The first image corresponds to the marked improvement in Fig. 3.4.1.

The distance matrix D_{ij} was calculated as described in detail in chapter 2: by comparing pixel-by-pixel, after pre-processing the images to adjust the alignment and to enhance the contrast. For the algorithms that don't use the distance matrix (OCSVM and distance to center of mass), the same pre-processing was used.

3.3.2. Face database

This publicly available database [36], kindly provided by AT&T Laboratories Cambridge, is constituted by face images (photographs of 40 subjects with 10 different images per subject) with outliers that were added similarly to Ju et al. [37]: first we rescaled the images to 64 by 64 pixels, and then, we added a square of noise to one randomly selected image per subject. Examples are shown in Fig. 3.3.2. When using the parameters proposed in Ju et al. [37] to generate the artifacts, all the methods have a perfect performance (average precision=1), so we generated the artifacts in the following manner: We used only square artifacts whose size we varied from 0 (no artifact added) to 64 (the whole image), the square was placed randomly in the image and its content was gray-scale pixels whose gray-scale value was randomly sampled such that the distribution was the same as the gray-scale value distribution of the combination of all the images in the database. We also generated a database with outliers whose brightness was modify by simply multiplying all the image by a constant factor.

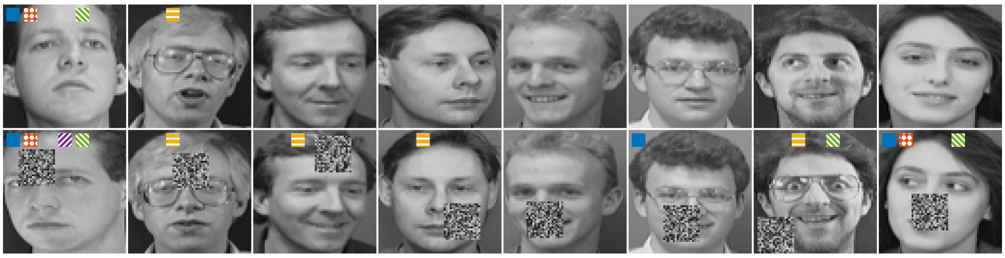


Figure 3.3.2.: Example images from the face database. Eight original images at the top, and eight images with added artifacts at the bottom. Marked images correspond to top 10% outlier score for OCSVM (Blue), distance to center of mass (Orange), IsoMap (Yellow), Percolation (Purple), and Ramaswamy (Green).

For this database (and also for the databases analyzed in appendix A, which also have added outliers), we generated two independent sets for each square size: one was used to find, in the case of the IsoMap and Ramaswamy methods, the optimal parameters, and the second one was used for testing.

For this database, the distance matrix was calculated as the Euclidean pixel-by-pixel distance.

3.3.3. Credit card transactions

This publicly available database [30–35] contains credit card transactions made in September 2013 by European cardholders. It contains 284807 transactions made in two days, of which 492 correspond to frauds. In order to preserve confidentiality, for each transaction the data set only includes the amount of money in the transaction, a relative time, and 28 features that are the output of a principal component analysis (PCA) of all the other metadata related to the transaction. In our analysis we divided the total dataset into 8 sets of about 4000 entries (due to computational constrains) according to the amount of the transaction and computed the distance as the euclidean distance using these 28 features.

3.4. Results

3.4.1. Anterior chamber OCT images

For the OCT database, there is no a priori definition of outliers (i.e., no ground truth), all the images were drawn from the same database. However, as a proxy for determining the performance of the outlier finding methods, we used the performance of the unsupervised methods proposed in chapter 2 when ignoring the images identified as outliers.

As removing outliers should improve the performance of machine learning algorithms, we performed two tests: first, we recalculated the correlation metrics presented in Table 2.1, removing the first n outliers that were identified by each method. Second, to test the significance of the improved performance, we repeated the calculation, now removing random images. The results presented in Fig. 3.4.1 confirm that removing the detected outliers improves the performance, while removing random images has no significant effect. We also see that IsoMap is the method that produces the highest improvement, while d2CM and OCSVM have low-significance performance improvement. For the IsoMap method we set the parameters to $d = 10$ and $k = 15$, while for the Ramaswamy method we used $k = 6$.

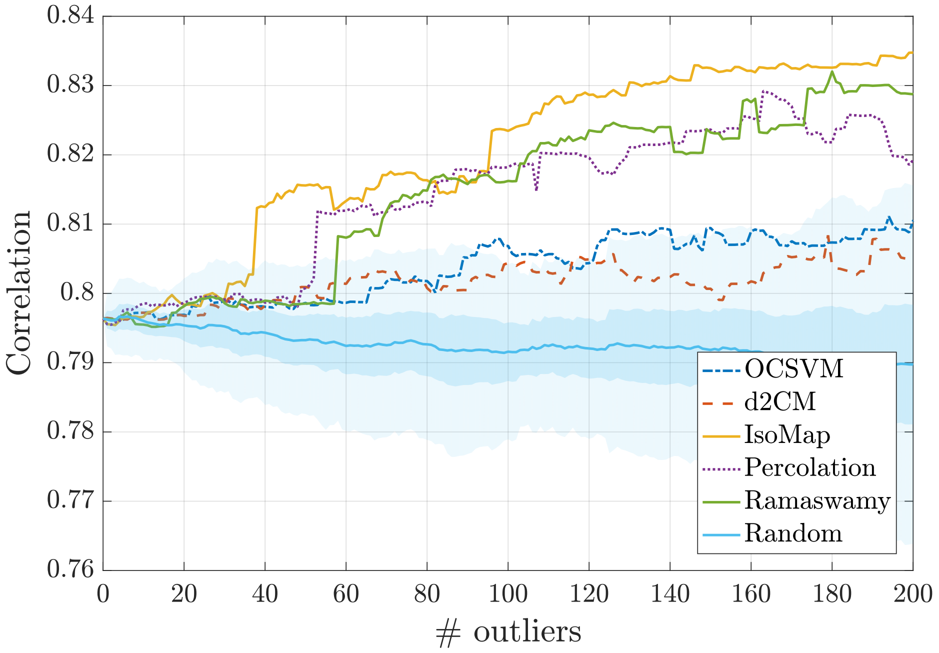


Figure 3.4.1.: Performance of the OCT image ordering algorithm as a function of the number of outliers that are removed from the database. As expected, we see that the performance, which is measured by the correlation coefficient between the feature returned by the ordering (unsupervised) algorithm and the feature provided by manual expert annotation (mean angle), improves as the outliers detected are removed. The different lines indicate the method of outlier identification and the colored region indicates results when the images removed are randomly selected, one standard deviation is shown in dark coloring, while three standard deviations is shown in light coloring. In this case, as expected, no significant change in the performance is seen. For some methods a sharp improvement is observed when eliminating one specific image (marked with a black circle), this image corresponds to the first one shown in Fig. 3.3.1.

3.4.2. Face database

For this database, as explained in Sec. 3.3.2, we generated artifacts artificially and tried to find the images presenting artifacts as outliers. We varied the size of the artifact generated to evaluate the robustness of the methods. For each size, we generated two different databases with artifacts (with the same parameters but different random seeds), we used the first one to optimize the parameters of IsoMap and Ramaswamy algorithms, and the second one to test the algorithms. We show the results of evaluating the performance on the second database for each square size in Fig. 3.4.2.(a), we used the average precision based on the precision-recall curve as performance measure, this measure computed as the area under the precision-recall curve [16] is more appropriate than other more commonly used metrics for class imbalance scenarios. In Fig. 3.4.2.(a) we see that Ramaswamy tends to slightly outperform all other methods, in particular, the percolation-based method shifts from being the worst method (when the squares are small) to the second best (when the squares are large). In Fig. 3.4.2.(d) we show the performance of the IsoMap method as a function of its parameters, we depict two zones with better performance, one with fairly low dimensionality and a low number of neighbors (more neighbors translate to a more linear mapping), and another zone with greater dimensionality and almost the maximum possible number of neighbors. In general, performance is very sensitive to parameter variations. In Fig. 3.4.2.(c) we show how altering the brightness of some images can also be perceived as outliers due to the distance measure used (Euclidean pixel-by-pixel).

Also, to evaluate how robust the methods are when changing the distance measure, we varied p in the Minkowski distance family (Eq. 3.2.1), and evaluated the methods for the parameters optimized for $p = 2$ (Euclidean), $p = 1$ and $p = 10$, the average precision as a function of p for the distance-based methods is shown in Fig. 3.4.2.(b). As we can see, for $p > 4$ Ramaswamy and Percolation-based perform similarly well, also, the parameters of Ramaswamy are very robust when changing p in the training set (the Ramaswamy method was also train with $p = 1$ and $p = 10$ obtaining the same parameters as for $p = 2$), while IsoMap is very sensitive to such changes.

We generated a different dataset whose outliers were images that, instead of having added noise, were multiplied by a constant (brightness) factor. We varied the brightness from 0 (the image being all black) to 3. The results of this study is shown in Fig. 3.4.2.(c).

3.4.3. Credit card transactions

In this database the ground truth (the fraud credit card transactions) is known and thus, the performance of the different methods is, as in the prior example, quantified with the average precision based on the precision-recall curve.

The database was divided into several subsets according to the amount of money of each transaction (see Fig. 3.4.3), each set (of around 4000 transactions) was further randomly divided into two sets in order to use one for training and the other one for testing. The results are summarized in Fig. 3.4.3 that displays the average precision for all testing sets. We can see that the performance of the methods is very heterogeneous.

To try to understand the origin of the large variability, we conducted an additional experiment in which we considered groups of 3900 normal transactions chosen at random (without considering the amount of the transaction) and 100 frauds also chosen at random, which were divided equally in training and test subsets. We repeat this experiment 8 times with different random seeds, and the results are presented in Fig. 3.4.4. in this experiment the average precision of the methods was increased due to a larger fraction of frauds in the test sets.

3.4.4. Discussion

Figure 3.4.5 presents the comparison of the results obtained with the five methods used, for the three databases analyzed. Figure 3.4.5.(a) summarizes the results for the OCT database, with

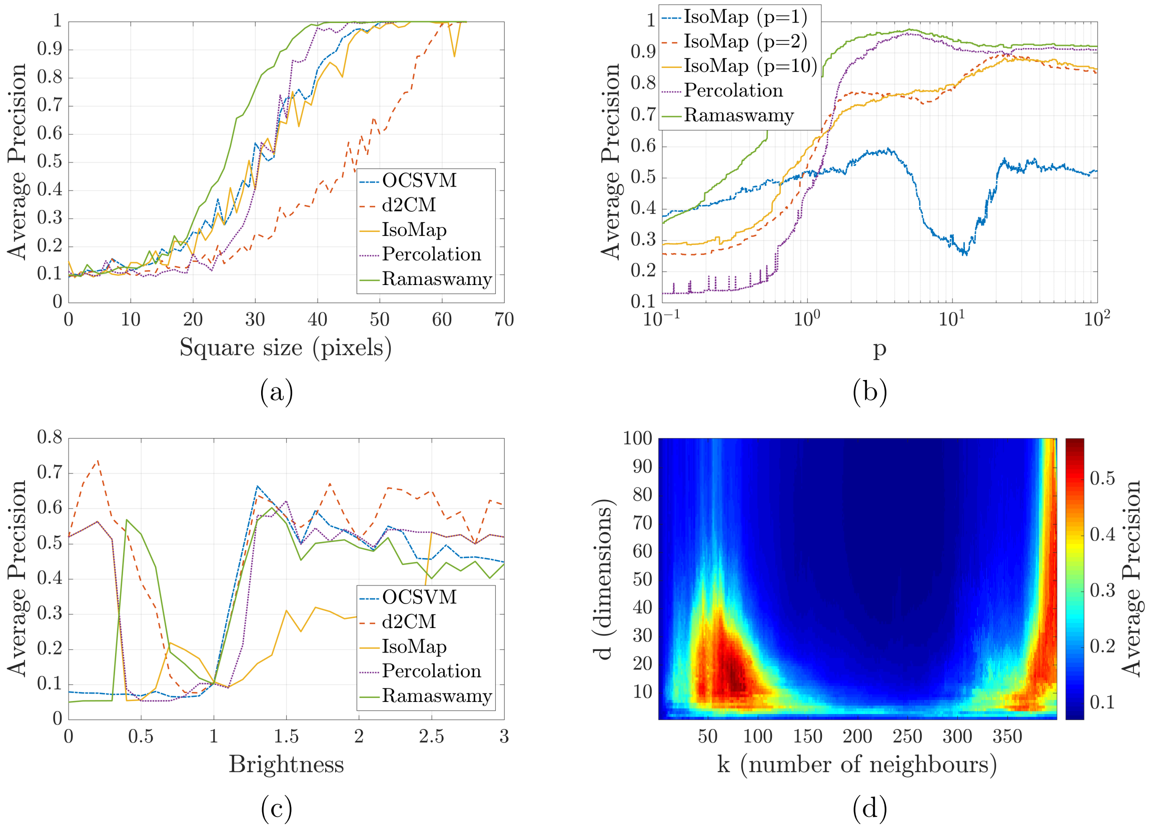


Figure 3.4.2.: Analysis of the face database. (a) Average precision as a function of the square size for the different outlier finding methods, One Class Support Vector Machine (OCSVM) in blue, distance to center of mass (d2CM) in orange, IsoMap in yellow, Percolation-based method in purple, and Ramaswamy in green. The Average precision was calculated on databases independently generated from those used to set the parameters of the methods. (b) Average precision as a function of the distance measure for different outlier finding methods with a square size of 36 pixels, IsoMap (trained with $p = 1$) in blue, IsoMap (trained with $p = 2$) in orange, IsoMap (trained with $p = 10$) in yellow, Percolation-based method in purple, and Ramaswamy in green. The Average precision was calculated on databases independently generated from those used to set the parameters of the methods. (c) Average precision as a function of the brightness multiplier of the outliers. (d) Average precision in the training set as a function of the IsoMap parameters with a square size of 30 pixels.

the boxplot we can see the minimum, first quartile, median, third quartile, and maximum of the correlation coefficient when varying the amount of outliers considered (corresponds to Fig. 3.4.1). Figure 3.4.5.(b) summarizes, in a similar manner, the results for the face database showing the boxplot of the average precision values when varying the square size [corresponds to Fig. 3.4.2.(a)]. Figure 3.4.5.(c) summarizes the results for the credit card transactions showing the boxplot of the average precision values when changing the amount range (corresponds to Fig. 3.4.3). As we can see in Fig. 3.4.5, the IsoMap and Percolation methods perform well in the three databases; their performance being either better than or comparable to the performance of the other three methods. Additional examples presented in appendix A confirm the good performance of IsoMap and Percolation methods.

Figure 3.4.2.(b) shows how the performance of distance-based methods is affected by the definition of the distance. We can see that the performance of all the methods depends on the definition of

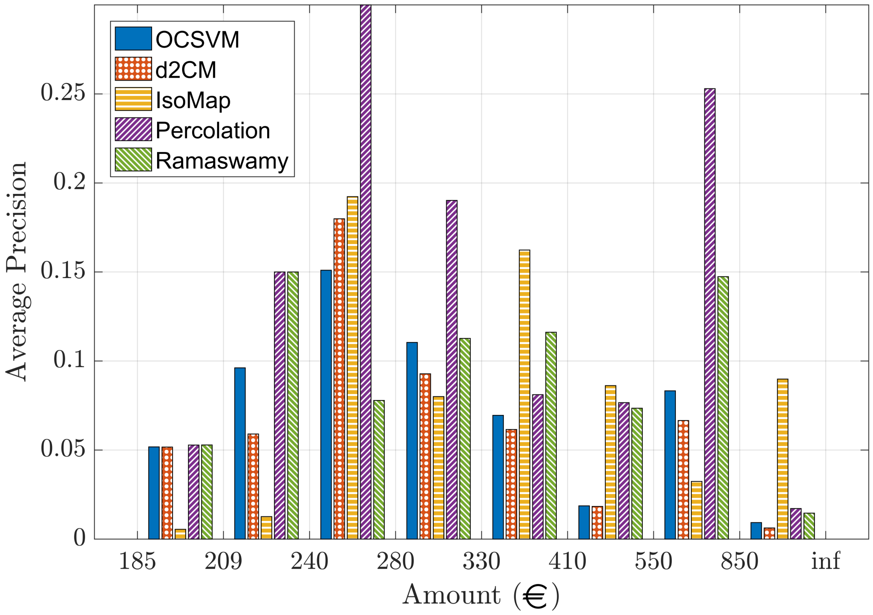


Figure 3.4.3.: Performance of all the outlier finding methods for the credit card transactions on the test subsets for each amount range.

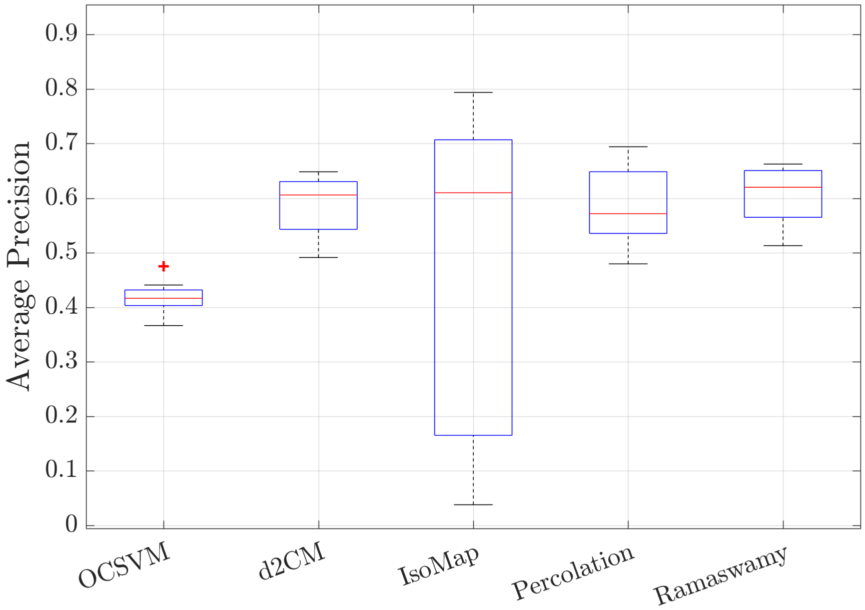
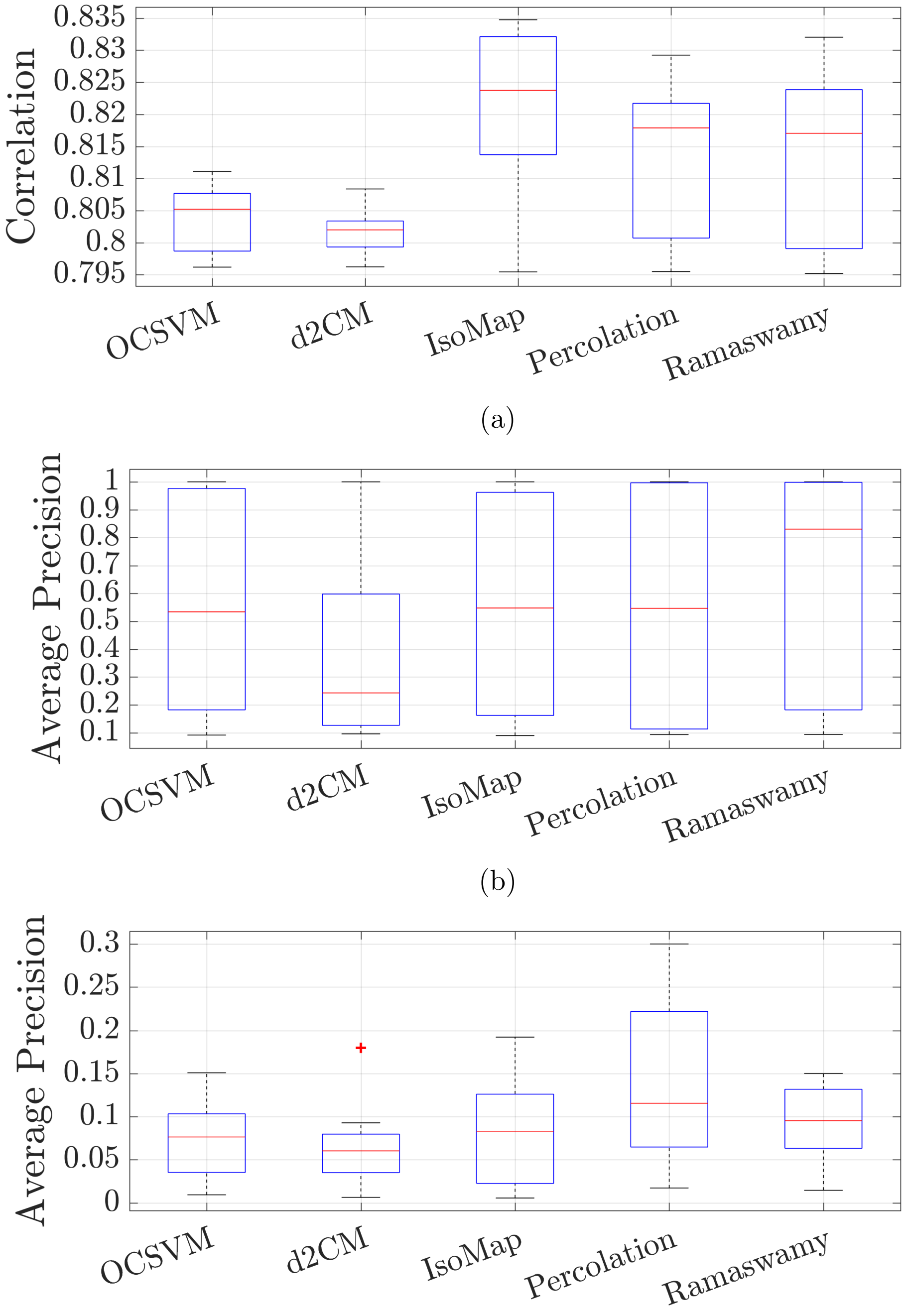


Figure 3.4.4.: Performance of all the outlier finding methods for the credit card transactions on the test subsets of the random groups. The random groups were generated by randomly choosing 3900 normal transactions and 100 frauds, and it was further randomly divided into two subsets, a training and a testing subsets.

the distance. The methods are also sensitive to changes in the preprocessing of the data, therefore, well-prepared data with a meaningful distance definition is needed for optimizing the performance of all methods.

It is important to consider how the two methods proposed here scale with the dimension of the data, d (i.e., the number of features of each sample), and the number of samples, N, in the database.



(a)

(b)

(c)

Figure 3.4.5.: Box-plot summarizing the results of all the outlier finding methods in all the databases. (a) Anterior Chamber OCT images, (b) Face database (only testing groups), (c) Credit Card transactions (only testing groups).

Since both methods begin by calculating the distance matrix, the processing time is at least of the order dN^2 because the calculation of the distance between pairs of elements linearly increases with d and quadratically with N. Both methods need to store in memory the distance matrix and analyze it, this imposes memory requirements that can limit their applicability for large datasets. In the case of IsoMap, this analysis is of order N^2. In the case of the percolation method, a threshold needs to be gradually varied in order to precisely identify the order in which the elements became disconnected from the giant component. This results in a runtime of the order of N^2 using the algorithm proposed in Newman and Ziff [118]. Regarding the dimensionality of the data, because both methods only need to hold in memory the N^2 distance matrix (and not the dN features of the N samples) they are suitable for very high dimensional data (where $d \gg N$) because once the distances of an element to all other elements have been computed, the d features of that element will not be needed again.

3.5. Conclusions

We have proposed two methods for outlier mining that rely on the definition of a meaningful measure of distance between pairs of elements in the dataset, one being fully unsupervised without the need of setting any parameters, and other which has 2 integer number parameters that can be set using a labeled training set. Both methods define a graph (whose nodes are the elements of the dataset, connected by links whose weights are the distances between the nodes) and analyze the structure of the graph. The first method is based on the percolation of the graph, while the second method uses the IsoMap nonlinear dimensionality reduction algorithm. We have tested the methods on several real and synthetic datasets (additional examples are presented in appendix A), and compared the performance of the proposed algorithms with the performance of a “naive” method (that calculates the distance to the center of mass) and two popular outlier finding methods, Ramaswamy and One Class Support Vector Machine (OCSVM).

Although the percolation algorithm performs comparably to (or slightly lower than) other methods, it has the great advantage of being parameter-free. In contrast, the IsoMap method has two parameters (natural numbers) that have to be selected appropriately. The performance of the methods varies with the dataset analyzed because the underlying assumption of what an outlier is, is different for the different methods. The percolation method assumes that the normal elements will be in one large cluster, with outliers being far from that cluster; IsoMap assumes that the normal elements lie on a manifold, and that outliers lie outside such manifold; the Ramaswamy and OCSVM methods assume that the outliers lie in a less densely populated sector of the space, while the “naive” method simply assumes that outliers are the furthest elements from the center of mass. These assumptions do not always hold, which results in the identification of normal elements as outliers. For example, in the OCT database there were some duplicated entries which were assigned by the Ramaswamy method the least outlier score, in spite of having a minor artifact.

The percolation algorithm is immune to duplicate entries, as it assigns the same outlier score as if there was only one element. On the other hand, the effect of duplicate entries on the IsoMap and “naive” methods is more difficult to asses, but is to be expected that if the duplicated elements are only few, they won’t have a large effect in the manifold learned, or in the center of mass calculated.

The execution time of both methods scales at least as dN^2 where d is the number of features of each item and N is the number of items in the database (as dN^2 is the time needed to compute the distance matrix). Therefore, the methods are suitable for the analysis of small to medium-size databases composed of high-dimensional items.

Chapter 4.

Network-based features for retinal fundus vessel structure analysis

Retinal fundus imaging is a non-invasive method that allows visualizing the structure of the blood vessels in the retina whose features may indicate the presence of diseases such as diabetic retinopathy (DR) and glaucoma. Here we present a novel method to analyze and quantify changes in the retinal blood vessel structure in patients diagnosed with glaucoma or with DR. First, we use an automatic unsupervised segmentation algorithm to extract a tree-like graph from the retina blood vessel structure. The nodes of the graph represent branching (bifurcation) points and endpoints, while the links represent vessel segments that connect the nodes. Then, we quantify structural differences between the graphs extracted from the groups of healthy and non-healthy patients. We also use fractal analysis to characterize the extracted graphs. Applying these techniques to three retina fundus image databases we find significant differences between the healthy and non-healthy groups (p-values lower than 0.005 or 0.001 depending on the method and on the database). The results are sensitive to the segmentation method (manual or automatic) and to the resolution of the images. The results presented in this chapter are published in Ref. [9].

4.1. Introduction

Fundus images are nowadays routinely used for the early diagnostic of ocular pathologies such as glaucoma [120–124] or diabetic retinopathy [125–133]. Other retinal imaging techniques are also used for manual and automatic diagnosis of these and other diseases [134]. Unsupervised algorithms can be used for automated retinal health screening, to differentiate normal fundus images from abnormal ones (age-related macular degeneration, diabetic retinopathy, and glaucoma) [135]. Studying the vascular structure of the retina can also advance our understanding of cardiovascular diseases [136, 137] and brain deceases, such as: Alzheimer [138] or dementia [139] due to changes in retinal microvasculature that may reflect similar changes in cerebral microvasculature. The performance of the analysis algorithms not only depends on the imaging technique and resolution [23] but also, on the methods used to segment the vessel network [140–148]. A main challenge for comparing the performance of different algorithms is that the performance of competitive algorithms is reaching the human intra-reader variability limit [129].

An analysis method with potential for diabetic retinopathy diagnosis is based on fractal analysis [149, 150]. While the fractal dimension of the blood vessels in the normal human retina is approximately 1.7 (consistent with a diffusion-limited growth process) [151], the fractal dimension of the vasculature tends to increase with the level of diabetic retinopathy [152]. However, the retinal fractal dimension varies considerably depending on the image quality, modality, and the technique used for measuring the fractal dimension [153]. The multifractal nature of the vascular network of the human retina [154, 155] and the reduction of the vasculature complexity with aging [156] have been reported. Fractal analysis has also been used to differentiate between healthy and pathological retinal texture [157].

Here we propose a new method that uses concepts inspired in network science [20, 158, 159]. We use the segmentation algorithm proposed in Ref. [41] to extract, from each digital fundus image, a tree-like graph where the nodes represent branching (bifurcation) points and endpoints, while the links represent vessel segments that connect two nodes. The graphs obtained are characterized by using the concept of node-distance distribution (NDD) [160], which is the fraction of nodes that are at distance d (shortest path) from a given node. We use as a reference node the optic disc (central node). To compare the extracted central distributions we use the Jensen-Shannon (JS) divergence that measures the distance between two probability distributions [69].

Precise graph comparison is a hard problem with many applications and different methods have been proposed in the literature (see Refs. [160, 161] and references therein). A main advantage of our approach is that it allows the comparison of graphs which have different numbers of nodes, and is appropriated for undirected and unlabelled graphs. Using a simpler metric (such as the Euclidean distance) can be more efficient for distinguishing different groups [162] but it only allows comparing graphs of the same size (i.e., with the same number of nodes), which is not the case for the graphs extracted from retina fundus images.

The proposed algorithm was tested on three databases of different size: a small high-resolution fundus (HRF) image database which comprises images of 15 patients with diabetic retinopathy, 15 with glaucoma and 15 without pathology; a large database, Messidor, where we used 230 images of patients with diabetic retinopathy classified in three groups, and 142 images of patients without pathology; and a medium size database from the Instituto de Microcirugía Ocular (IMO) which contains 70 images of glaucoma patients, and 23 images of patients without pathology. By means of nonlinear dimensionality reduction techniques we show that the DR, glaucoma and healthy groups, have statistically significant different features. To support these results, we also calculate the fractal dimension of the images (segmented and skeletonized versions) and find significant differences between the three groups, which are fully consistent with the results of the graph dissimilarity analysis.

4.2. Methods

In this section we present the algorithms proposed for unsupervisedly retrieve features from images in a database. We also present the three databases we used to test our algorithms.

All the methods make use of the result of an unsupervised segmentation algorithm that was adapted from the one proposed in Ref. [41]. We start by filtering the photograph to enhance the contrast between the vessels and the background, and then we perform the Graph-based segmentation algorithm as proposed in Ref. [41], further details on the segmentation process can be found in the appendix B. We refer to the raw segmentation result as a binary image whose pixels are 1 if they belong to a vessel and 0 if they belong to the background.

4.2.1. Fractal dimension

The box counting algorithm is a well-known method for estimating the fractal dimension (FD) of a geometrical object [163]. It is based on the covering of an object with a grid of boxes of size ε and counting the number of boxes with information inside, $N(\varepsilon)$. The dimension is an exponent that quantifies the scaling of $N(\varepsilon)$ with the size ε as $\varepsilon \to 0$, $N(\varepsilon) \sim \varepsilon^{-D}$, where D is the fractal dimension which can be cast as an equation:

$$D = \lim_{\varepsilon \to 0} \frac{\log(N(\varepsilon))}{\log(1/\varepsilon)} \tag{4.2.1}$$

where $\varepsilon \to 0$ is used to ensure coordinate invariance. We apply the box counting method to both, the raw segmented image (i.e., a binary image that includes all the pixels that correspond to vessels); and to the skeletonized image (i.e., a binary image where the width of each vessel segment was reduced to one pixel, without changing the length, location and orientation of each segment).

4.2.2. Graphs extracted from segmented images

With the information retrieved from the segmented images (raw and skeletonized) we construct undirected graphs where the nodes represent the branching (bifurcation) points and the endpoints, and the links represent vessel segments that connect pairs of nodes. The links have associated weights that represent the cost of transporting matter from one node to the other. If nodes i and j are not connected, $w_{i,j} = 0$, while if there is a segment connecting them, $w_{i,j} \neq 0$. In order to test different possibilities using the values of the length, $L_{i,j}$, and the width, $W_{i,j}$, of the segment that connects nodes i and j, the weight of the link is defined as:

$$w_{i,j} = (L_{i,j})^l \, (W_{i,j})^a \tag{4.2.2}$$

being l and a adjustable exponents, exploring, in this way, the group classification in terms of the length, the width and any product of powers of these two features. The length $L_{i,j}$ and the width $W_{i,j}$ of each link can be computed using the information contained in the skeletonized and raw segmentations. The length accounts for the number of pixels spanned by each link in the skeletonized version while the width can be estimated from the number of pixels ($N_{i,j}$) each link has in the raw segmented mask as $N_{i,j} = L_{i,j} \times W_{i,j}$.

4.2.3. Network measures

Structural differences between the extracted graphs were characterized by using the measures described in this section, which provide probability distribution functions (PDFs) that can be mutually compared by using nonlinear dimensionality reduction (NLDR) techniques.

4.2.3.1. Distributions of distances to the central node

The node distance distribution (NDD) measures the heterogeneity of a graph in terms of the connectivity distances, and allows the precise comparison of two graphs, by quantifying the differences between distance-based PDFs extracted from the graphs. It is based on computing, for each node i, the probability that another node j is connected to i with a path of distance d.

To apply the NDD concept to the tree-like graphs extracted from the segmented images, we consider only the distribution of distances to the central node that represents the optic nerve (because all the transported blood comes from and returns to this node). Thus, we analyze the Central NDD (C-NDD) PDF that gives the distribution of distances of the nodes to the central one. The distance of one node to the central one is defined as the sum of the weights of the shortest path.

As an example, in Fig. 4.2.1, the distance of the selected node to the central one is the sum of the weights of the three links that connect the two nodes. The distribution $P_{CNDD}\left(d\right)$ is the fraction of nodes whose weighted distance from the central node is d, i.e., that the weighted shortest path of these nodes to the central node consists of links whose weights add up to d.

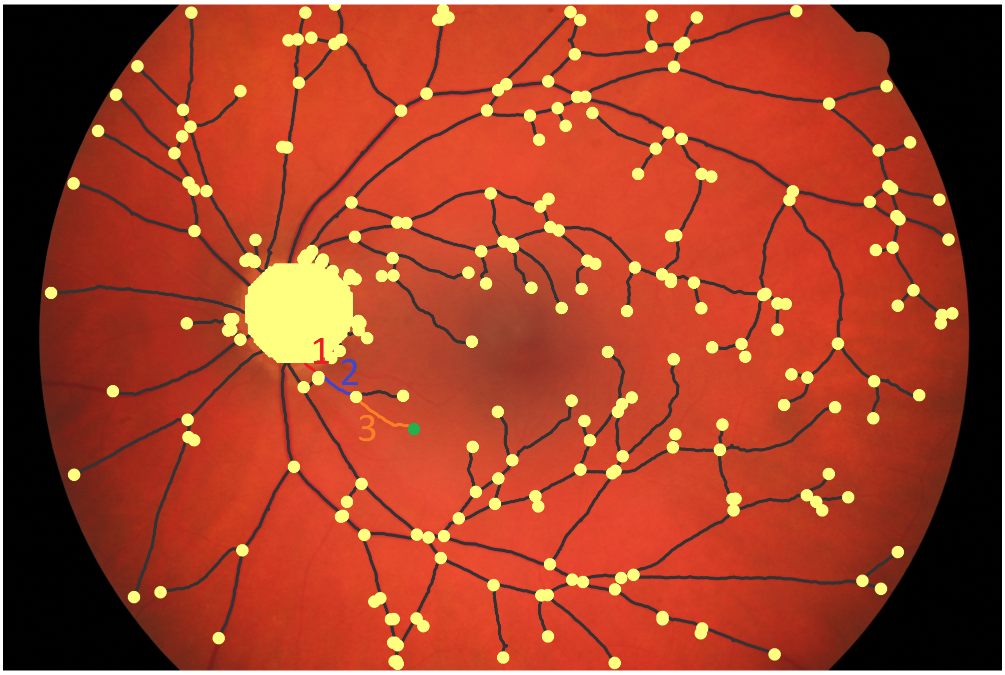


Figure 4.2.1.: Example result of the automatic segmentation algorithm on top of the original image, nodes are shown in light yellow, while links are shown in dark gray. An example of a shortest path from the optical disk to the node highlighted in green is shown, it consists of three links each one having its own weight according to Eq. 4.2.2.

A variation of the Central NDD is the central mean weight distribution (CMWD), which is the distribution of average weights, i.e., the sum of the weights of the links that connect two nodes, divided by the number of links.

4.2.3.2. Weighted degree distribution

The degree distribution, $P_{DD}\left(k\right)$, is a popular measure to describe the heterogeneity of the nodes of a graph. $P_{DD}\left(k\right)$ is just the probability that a node has k links. In regular graphs all the nodes have the same number of links, and therefore, $P_{DD}\left(k\right)$ is the delta-distribution, while in random graphs, $P_{DD}\left(k\right)$ has a Gaussian-like shape. In weighted graphs, $P_{WDD}\left(s\right)$ is the distribution of the strengths of the nodes (the strength, s, of a node i is the sum of the weights of its links, i.e., $s_i = \sum_j w_{i,j}$).

4.2.3.3. Unsupervised nonlinear dimensionality reduction

The analyses described above provide us, for each image, with various probability distributions (one for each combination of l, a). These multidimensional descriptors carry several features which are often redundant. By using a nonlinear dimensionality reduction technique (NLDR), we are able to represent each distribution as a single point in a two-dimensional plane. In order to do that, we compare the distributions in a pair-wise manner using the Jensen-Shanon (JS) divergence [69]. In this way, we obtain a matrix, P, of dimension $N \times N$ (N being the total number of images analyzed) whose elements $p_{i,j}$ are the distance (JS divergence) between the probability distributions extracted from images i and j. Then, using this matrix P as input for the IsoMap algorithm [11], it returns two features that are the coordinates of a point in a plane. This plane captures similarities and differences in the distributions such that similar distributions are represented as points close together and different distributions are represented as points far away from each other. It is worth noting that the algorithm is fully unsupervised, i.e., no prior image information (diagnosis) is needed at any step.

4.3. Data

We used three different databases to test our algorithms.

4.3.1. High-Resolution Fundus (HRF) Image Database

The HRF is a public database[1] [28] which contains 45 color fundus images divided in 15 images of healthy patients, 15 images of patients with diabetic retinopathy and 15 images of glaucomatous patients. These images were captured using a Canon CR-1 fundus camera with a field of view of 60 ° and have a size of 3504 × 2336 pixels.

This database also includes a manual segmentation of the vessel network performed by a human expert. For comparison purposes we have also analyzed this set of images, as well as the images resulted from our automated segmentation method described in the appendix B.

4.3.2. Messidor Image Database

This database, kindly provided by the Messidor program partners[2] consists of 1200 color fundus images taken with a field of view of 45º and resolutions ranging from 1440 × 960 pixels to 2304 × 1536 pixels [29]. Each image is categorized in one of four groups, corresponding to a diabetic patients without diabetic retinopathy and three increasing stages of diabetic retinopathy.

As our method is sensitive to changes in the images resolution, we worked with the first 400 images in the database that have a resolution of 2240 × 1488 pixels. Out of the 400 images we discarded 28 images in which the algorithm either failed to segment the network or to find the optic nerve and analyzed 372 images: 230 with diabetic retinopathy and 142 without.

[1] Download available at: https://www5.cs.fau.de/research/data/fundus-images/
[2] see http://www.adcis.net/en/DownloadThirdParty/Messidor.html

4.3.3. Instituto de Microcirugía Ocular (IMO) Images

We also analyzed 93 images from patients at IMO (Ocular Microsurgery Institute [68]): 23 from healthy subjects and 70 from glaucoma patients. The images have a field of view of 45º and a resolution of 2000 × 1312 pixels. We used images from IMO database from consenting patients. The use of this data was approved by IMO's Ethical committee for clinical study with date October 9th, 2018.

4.4. Results

We applied the analysis tools described in section 4.2 to the three retina fundus image databases. For the HRF database we performed the analysis using our automated segmentation and the manual segmentation provided with the database.

The algorithms were implemented in MatLab (segmentation, network retrieval, IsoMap) and Python (fractal analysis, network analysis) and their runtime using personal laptops was between 5 and 35 seconds per image depending on the resolution. These runtimes could be improved by rewriting the algorithms in a compiled language, however, they provide a rough assessment of the complexity of the algorithms.

We have summarized the results in two tables (4.1 and 4.2) where p-values were used to assess the statistical significance of the results obtained with the different methods. They were calculated with a t-test (using MatLab function *ttest2*) of the null hypothesis that the two samples come from distributions with equal means. For comparison, we also include references to other papers that have analyzed these databases and provided p-values.

Analysis	l	a	MESSIDOR p-Val.	HRF Automated p-Val.	HRF Manual p-Val.
C-NDD	1	-2	**0.011**	**0.0048**	**7.1e-05**
C-NDD	1	2	0.29	0.57	**6.4e-09**
CMWD	1	-2	0.82	0.68	**1.0e-10**
WDD	0	1	**0.0028**	**0.0070**	**8.0e-15**
Nodes	-	-	**0.0052**	0.074	0.69
Links	-	-	**0.0066**	0.073	0.99
Endpoints	-	-	**0.0054**	0.070	0.29
Bifurcation points	-	-	**0.0050**	0.082	0.65
FD skeletonized	-	-	0.23	0.88	0.68
FD raw	-	-	**1.3e-06**	**0.0096**	**0.0026**
FD best direction	-	-	**9.0e-12**	**4.5e-05**	**8.5e-06**
Best result using FD proposed in Ref. [153]	-	-	**0.01**	-	-
FD result in Ref. [164]	-	-	**0.0088**	-	-

Table 4.1.: p-values obtained by comparing the features extracted from the Messidor and HRF databases (automated and manual segmentations) of the groups with and without diabetic retinopathy (p-values smaller than 0.05 in **bold**).

Analysis	l	a	IMO p-Val.	HRF Automated p-Val.	HRF Manual p-Val.
C-NDD	1	-2	0.27	**0.0037**	**1.5e-06**
C-NDD	1	2	**5.8e-05**	**6.7e-05**	**0.00012**
CMWD	1	-2	**0.0066**	**0.00015**	**7.6e-07**
WDD	0	1	0.99	0.087	**1.1e-15**
Nodes	-	-	**0.012**	**0.0041**	**0.029**
Links	-	-	**0.012**	**0.0046**	**0.012**
Endpoints	-	-	**0.015**	**0.0042**	0.10
Bifurcation points	-	-	**0.0089**	**0.0034**	**0.026**
FD skeletonized	-	-	**0.0028**	**0.0038**	0.057
FD raw	-	-	0.11	**0.00058**	**0.0027**
FD best direction	-	-	**0.0015**	**0.00041**	**9.6e-08**
Best result in Ref. [165]	-	-	-	**<1e-6**	-

Table 4.2.: p-values obtained by comparing the features extracted from the IMO and HRF databases (automated and manual segmentations) of the healthy and glaucoma groups (p-values smaller than 0.05 in **bold**).

4.4.1. Fractal dimensions

Figure 4.4.1 displays, for the HRF database, the fractal dimension of the raw segmented images *vs.* the fractal dimension of the skeletonized ones for both the automated (panel a) and manual (panel b) segmentations. In the scatter plots each point corresponds to an image, while the ellipsoids represent the square root of the covariance matrix of each group (Diabetic, Glaucoma, and Healthy).

In both segmentations a clear distinction between healthy and non-healthy groups is obtained. In addition, with the automated segmentation a clear segregation between the three groups is obtained (panel a), which is not seen in the analysis of the manual segmentation (panel b).

Similar segregation between healthy and non-healthy groups is obtained for the Messidor and for the IMO databases, with p-values of the order of 9e-12 for Messidor (see Table 4.1) and 0.0015 for IMO (see Table 4.2).

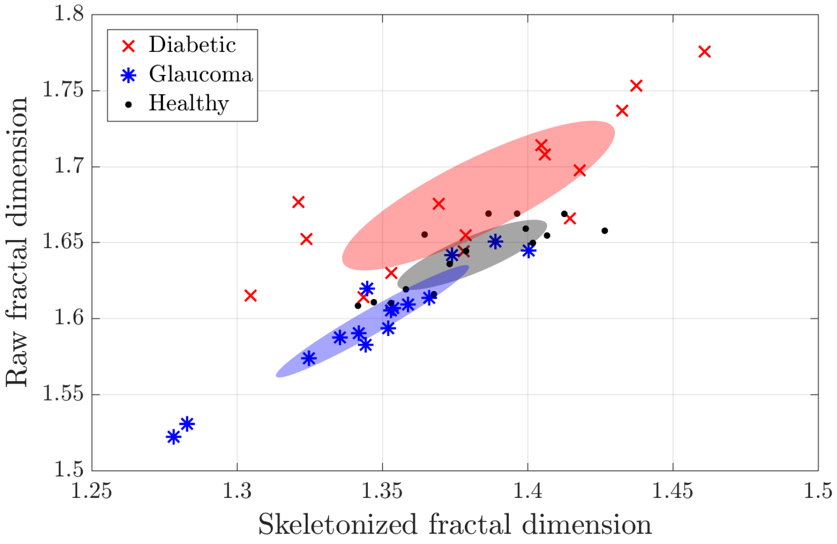
4.4.2. Central NDD

Figure 4.4.2 displays, for the HRF database, the results obtained from the C-NDD analysis with $l = 1$, $a = -2$ (since they provide the best differentiation between groups), for the automated (panel a), and manual (panel b) segmentations. As in Fig. 4.4.1, here each point corresponds to the two features coming from the NLDR technique applied to the C-NDD histogram of each image and ellipsoids represent the square root of the covariance matrix for each group.

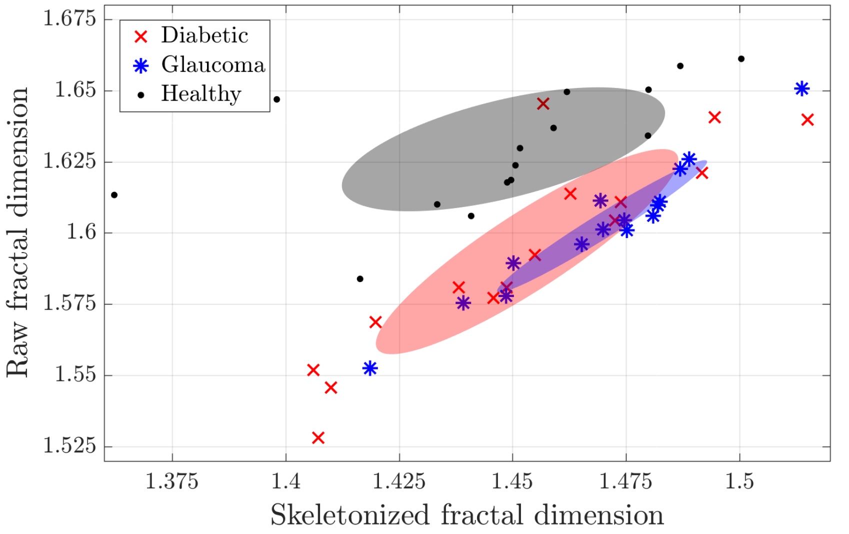
Again a clear distinction between the groups is obtained, with p-values (see Tables 4.1 and 4.2) in the order of 0.005 for automated segmentation and 1e-15 for manual segmentation.

4.4.3. Central mean weight distribution

Figure 4.4.3 displays, for the HRF database, the results obtained from the CMWD analysis with $l = 1$, $a = -2$, using the automated (panel a), and manual (panel b) segmentations. Figure 4.4.4 shows the corresponding histograms. Here we see that for both segmentations, the distributions corresponding to healthy subjects tend to be more skewed to the left with respect to the



(a)



(b)

Figure 4.4.1.: Fractal dimension analysis of the HRF database using the (a) automated and (b) manual segmentation. In both plots the horizontal axis denotes the fractal dimension of the skeletonized mask while the vertical axis accounts for the fractal dimension of the raw segmented mask. Each point represents the fractal dimensions of one image, while the ellipses represent the square root of the covariance matrix of each group. In (a) we note that the three groups are well separated (p-values 4.5e-05, 0.00041), while in (b) the healthy group is well separated from the non-healthy ones (p-values 8.5e-06, 9.6e-08). In both plots we note that using the two fractal dimensions improves the separation, in comparison to using only one.

pathological ones, however, as shown in Fig. 4.4.3.(a), the diabetic group and the normal group are indistinguishable using the automatic segmentation.

Similarly to the previous C-NDD analysis, in the manual segmentation the two non-healthy groups are clearly separated from the healthy one, while on the automated segmentation only the glaucoma group is found to be statistically different from the healthy one. The same results (not shown) hold for the Messidor and IMO databases.

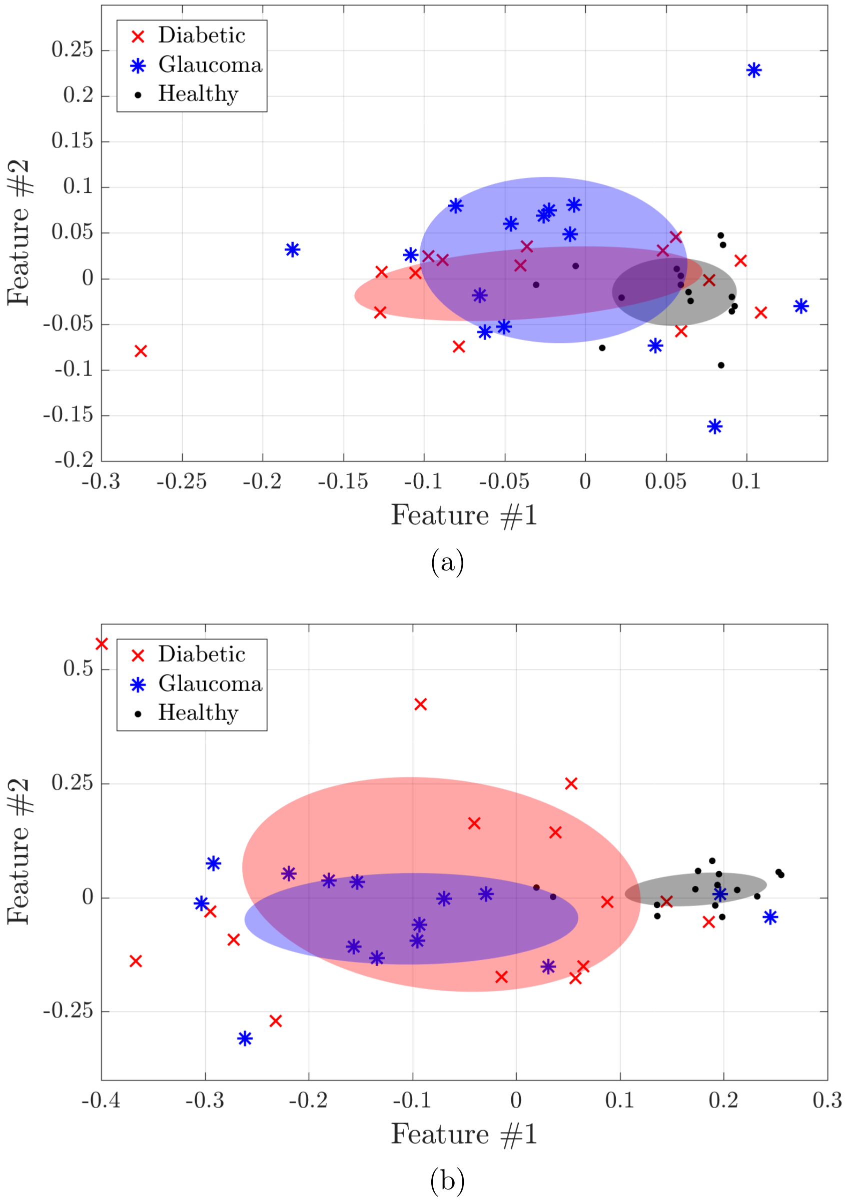


Figure 4.4.2.: The panels display the IsoMap features extracted from the HRF database, using the automated (a) and manual (b) segmentations, with C-NDD analysis. The weights (Eq. 4.2.2) are defined with $l = 1$ and $a = -2$. Here we note that the C-NDD analysis perform very well in the manual segmentation (giving a clear distinction between the healthy and unhealthy groups), while in the automated segmentation it doesn't provide a clear separation.

Comparing C-NDD and C-MWD, one can see that the prior performs better for diabetic retinopathy while the latter performs better for glaucoma. The p-values are summarized in Tables 4.1 and 4.2.

4.4.4. Weighted degree distribution

Figure 4.4.5 displays, for the HRF database, the results obtained from the WDD analysis with $l = 0$, $a = 1$ (i.e., the weight of the link just accounts for the vessel width), using the automated (panel a) and manual (panel b) segmentations.

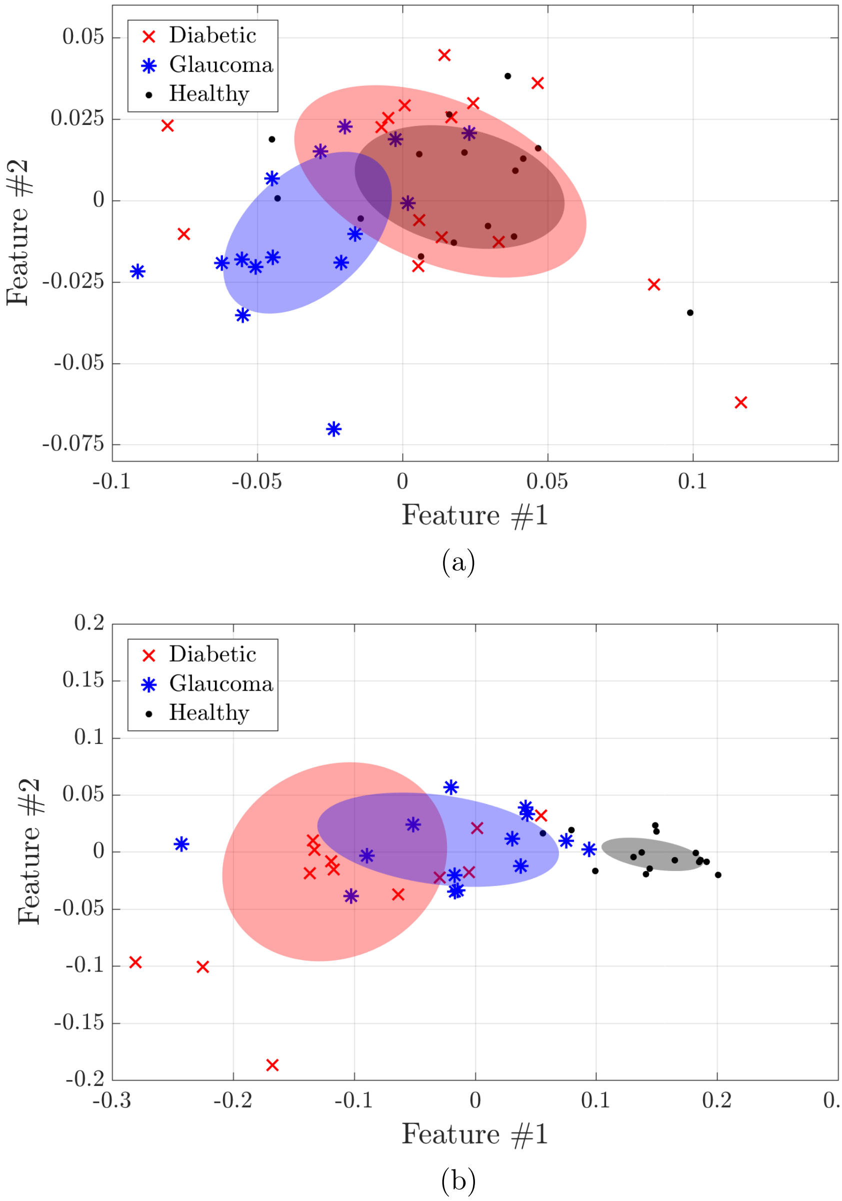


Figure 4.4.3.: The panels display the IsoMap features extracted from the HRF database, using the automated (a) and manual (b) segmentations, with C-MWD analysis. The weights (Eq. 4.2.2) are defined with $l = 1$ and $a = -2$. Here we note that the C-MWD analysis perform very well in the manual segmentation (giving a clear distinction between the healthy and unhealthy groups), while in the automated segmentation it doesn't provide a clear separation.

Here we observe (as in the previous analysis, compare Fig. 4.4.3) that in the manual segmentation the two non-healthy groups are quite separated from the healthy one. In the automated segmentation instead, only the diabetic group is statistically different from the healthy one. The same results hold for the Messidor database (see appendix C), while for the IMO database (not shown) the results are not significant. p-values are summarized in Tables 4.1 and 4.2.

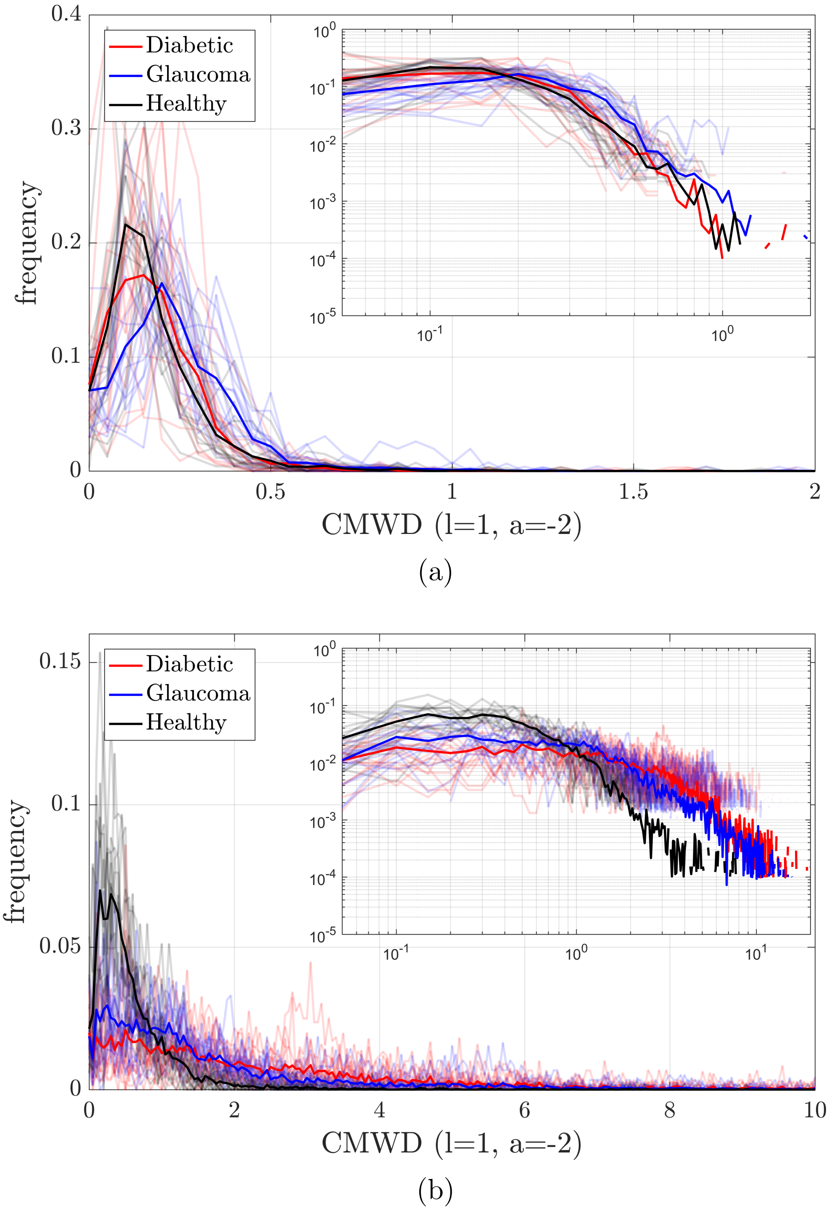
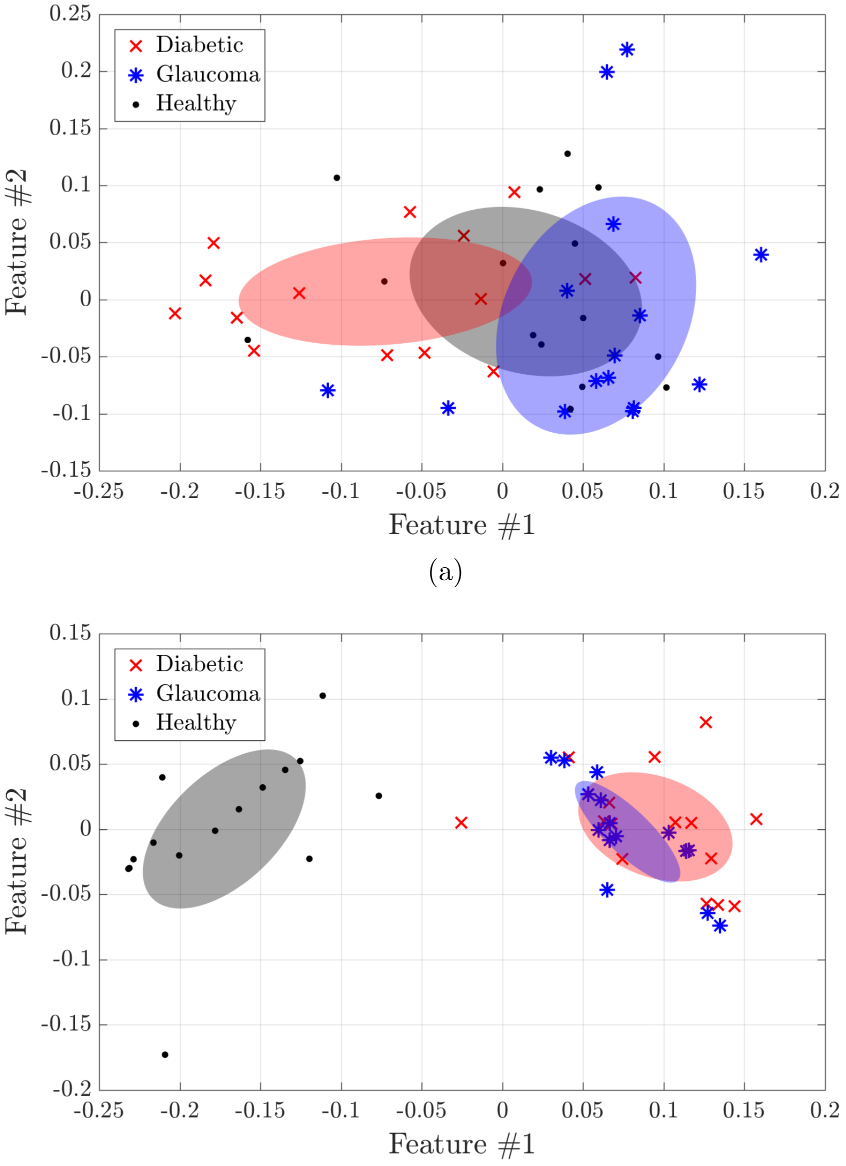


Figure 4.4.4.: The panels display the raw histograms of the C-MWD extracted from the HRF database (with logarithmic scale in the insets), using the automated (a) and manual (b) segmentations (it corresponds to Fig. 4.4.3). The weights (Eq. 4.2.2) are defined as $l = 1$ and $a = -2$. The mean histogram of each group is shown in a solid line, while every individual histogram was plotted with semi-transparent lines using the color corresponding to its group.

4.4.5. Analysis of other network features

We also analyzed other network features, such as the number of links, the number of nodes, the number of endpoints (nodes with only one neighbor) and the bifurcation points (nodes with 3 or more neighbors). The results, also presented in Tables 4.1 and 4.2, suggest that these basic features are not as informative as the features presented before: they have less significance (in terms of p-values) for identifying statistical differences between groups.



(a)

(b)

Figure 4.4.5.: IsoMap features obtained from the (a) automated and (b) manual segmentations, with WDD analysis. The weights (Eq. 4.2.2) are defined with $l = 0$ and $a = -1$. We observe that with the manual segmentation the healthy group is clearly separated from the non-healthy ones, while with the automated segmentation, the three groups are different, but they are not fully separated.

4.5. Discussion

In Table 4.1 we show the results of the analysis of the diabetic and healthy groups from HRF and Messidor databases and in Table 4.2 the analysis of the Glaucoma and the healthy groups from HRF and IMO databases. In almost every case, the algorithms performed better when using the HRF manual segmentation compared to the automatic segmentation. We think that this is due to the intrinsic better quality of the manual segmentation, although it may also be attributed to observer bias. We don't have a clear explanation of why, in a few cases, the automatic segmentation performs better than the manual one.

For the diabetic condition, in both Messidor and HRF with automated segmentation the best analysis turned out to be the best direction of the fractal dimension plane (i.e., a linear combination of both proposed fractal dimensions) while for the manual segmentation the best analysis was WDD with $l = 0$ and $a = 1$ (which was also the second best for Messidor), although all the proposed network based analysis were statistically significant. The methods that performed consistently well with both databases (and both segmentations) regarding diabetic retinopathy were C-NDD ($l = 1$, $a = -2$), WDD ($l = 0$, $a = 1$), the fractal dimension of the raw segmented image, and the best direction in the fractal dimension plane.

For the glaucoma case, in both IMO and HRF with automated segmentation the best results were obtained using the C-NDD with $l = 1$ and $a = 2$, while for the manual segmentation of HRF the best analysis was, again, WDD with $l = 0$ and $a = 1$, although all the proposed network based analysis were statistically significant. The methods that performed consistently good with both databases (and both segmentations) regarding glaucoma were C-NDD ($l = 1$, $a = 2$), CMWD ($l = 1$, $a = -2$), number of nodes, number of links, number of bifurcation points, and the best direction in the fractal dimension plane.

The parameters l and a chosen to test the network analysis have all a clear physical interpretation. The set $l = 0$ and $a = 1$ is simply the width of the corresponding vessel. The set $l = 1$ and $a = 2$ is proportional to the volume of such vessel. And finally, the set $l = 1$ and $a = -2$ can be related to the flow resistance of the vessel. Some other sets were tested such as the ones corresponding to the length and cross section of the vessel, obtaining no significant results.

It should be noted that the analyzed network is a 2-dimensional projection of the real 3-dimensional retina network, this implies that there are some nodes in it which, in reality, correspond to crossovers of veins and arteries. This alters the extracted features in two ways, by generating spurious nodes whose links are fictional, and by generating spurious shortest paths to the optical nerve. The problem of distinguishing arteries from veins in fundus photographies is highly non-trivial [166–168]. We have tested an algorithm based on prior work [169,170] that eliminates spurious nodes (for example, those that have 4 links), but we found that the modification did not improve the performance of the proposed measures, while it added complexity and more parameters to the algorithm. We speculate that there are two reasons why the pruning of the spurious nodes does not improve the performance: 1) because most measures rely on the shortest path to the optical nerve, and only few paths are modified when comparing the "true network" with the 2D projected one, and 2) because the same kind of artifacts are present in all the images, thus, the comparison between images is still fair.

Our findings are consistent with the results recently reported in Ref. [171], where topological data analysis (TDA) was applied to the fundus images of diabetic retinopathy patients and healthy subjects in the HRF and MESSIDOR databases. The TDA features (that characterize connected components and holes in the images) allowed to discriminate between healthy patients and those with diabetic retinopathy in the HRF database but not in the MESSIDOR database, a fact that was interpreted as a due to the much lower resolution of MESSIDOR.

4.6. Conclusion

We have demonstrated that the network-based features extracted from fundus images are useful for detecting topological changes produced in patients with diabetic retinopathy and glaucoma.

For both diseases, the proposed network features we have proposed are able to separate the healthy group and the unhealthy groups with extremely high statistical significance. We have also compared our results with those obtained from fractal geometry analysis, and we have shown that using both fractal dimensions (raw segmented, and skeletonized) improves the separation between the groups, in comparison to using only one. The most statistically significant results were obtained using high resolution images (the HRF database), and in particular, when using the manual segmentation provided with the database. We found that analyzing the manual segmentation of the HRF database with the weighted degree distribution (see Fig. 4.4.5) perfect classification could be achieved for both studied pathologies, and we note that this is not the case when using fractal analysis. In our study, it is apparent when comparing the results of manual and automated segmentation (in both diseases) that with the manual segmentation the classification performs almost always better than with the automated segmentation. Thus, improving the segmentation algorithm would probably improve the performance of the features derived from it. When analyzing images with lower resolution, the results show that the differences among the groups are not as statistically significant, and thus, we conclude that the topological differences found correspond to differences in the thinnest vessels of the network.

When analyzing diabetic patients, the weights that performed the best were the widths ($l = 0$ and $a = -1$ in Eq. 4.2.2) and length/ $(\text{width})^2$ ($l = 1$ and $a = -2$ in Eq. 4.2.2). This can be understood by considering that diabetic retinopathy causes neovascularization that consists of thin vessels, and can also affect the vessel flow capacity. When analyzing glaucoma patients, the weight that performed the best was the volume ($\propto$ length $(\text{width})^2$). This can be understood by considering that glaucoma is linked to an increase of the intraocular pressure, which can increase the volume of the vessels.

The measures proposed in this chapter demonstrated very good performance in retina fundus images of different resolution, and of patients with different diseases. Therefore, it will be interesting to explore their potential with other vascular-related diseases.

Chapter 5.

Machine learning algorithms for predicting the amplitude of chaotic laser pulses

Forecasting the dynamics of chaotic systems from the analysis of their output signals is a challenging problem with applications in most fields of modern science. In this work, we use a laser model to compare the performance of several machine learning algorithms for forecasting the amplitude of upcoming emitted chaotic pulses. We simulate the dynamics of an optically injected semiconductor laser that presents a rich variety of dynamical regimes when changing the parameters. We focus on a particular dynamical regime that can show ultra-high intensity pulses, reminiscent of rogue waves. We compare the goodness of the forecast for several popular methods in machine learning, namely deep learning, support vector machine, nearest neighbors and reservoir computing. Finally, we analyze how their performance for predicting the height of the next optical pulse depends on the amount of noise and the length of the time-series used for training. The results presented in this chapter are published in Ref. [10].

5.1. Introduction

Optically injected semiconductor lasers have a rich variety of dynamical regimes, including stable locked emission, regular pulsing and chaotic behavior [38, 172]. These regimes have found several practical applications. For example, under stable emission the laser emits light at the injected wavelength (the so-called injection-locking region) and has a high resonance frequency and a large modulation bandwidth [173], which have broad applications for optical communications. The regular pulsing regime can be used for microwave generation [174, 175], while the broad-band chaotic signal can be exploited for ultra-fast random number generation [176].

In turn, the output of the laser in the chaotic regime can be used for testing new methods for data analysis, and in particular, for time series prediction. Predicting the dynamical evolution of complex systems from the analysis of their output signals is an important problem in nonlinear science [177–179], with a wide range of interdisciplinary applications. In these "big data" days, a significant number of researchers are focusing on developing novel methods for time series forecasting based on machine learning algorithms [180–184].

Delay embedding and recurrent neural networks have been used to predict the evolution of chaotic systems such as the Lorenz system and the Mackey-Glass system [185]. Locally linear neurofuzzy models [186] and support vector machine [187] have also been used to forecast chaotic signals. Here, in contrast with previous works, we do not attempt to forecast the evolution of a chaotic system, but the amplitude of the next peak in the observed signal.

As a case study, we consider the dynamics of an optically injected laser. We simulate the laser dynamics using a well-known rate equation model [38,39], and use the chaotic regime to compare the performance of several machine learning algorithms (deep learning, support vector machine, nearest neighbors and reservoir computing) for forecasting the amplitude of the next intensity pulse.

Our main motivation to study this system is that it can be implemented experimentally and we hope that our work will motivate the analysis of real data. An important characteristic of this laser system is that it has control parameters (that can be varied in the experiment) that allow to generate time series with or without extreme pulses.

Therefore, in the simulations, within the chaotic regime, we consider two different situations: the intensity pulses display occasional extreme values (so-called optical rogue waves [79, 188]) or the intensity pulses are irregular but do not display extreme fluctuations. In the first case, the probability distribution function (pdf) of pulse amplitudes is long tailed, while in the second case, it has a well-defined cut off.

The possibility of predicting and suppressing extreme pulses in a chaotic system has been demonstrated in Ref. [189], but in this work the authors did not attempt to predict the pulse amplitude but rather the occurrence of a very high pulse whenever the trajectory approached a particular region of the phase space. To shed light on the limits of the forecast of extreme events, we consider dynamical regimes with and without extreme pulses, produced by the same underlying system, and we attempt to predict the amplitude of the next pulse, regardless of whether it is normal or extreme. In our system we find that, while both regular and extreme pulses can be forecasted, the existence of extreme pulses bounds the prediction accuracy. In an experimental setup, observational noise and the limited bandwidth of the detection system (photodiode, oscilloscope) can further limit the predictability of the pulse amplitude.

5.2. Model

We simulated the dynamics of the complex optical field E and the carrier population N in a semiconductor laser with optical injection using the following rate equations [172, 190]:

$$\frac{dE}{dt} = \kappa\left(1 + i\alpha\right)\left(N - 1\right)E + i\Delta\omega E + \sqrt{P_{inj}} + \sqrt{D}\xi\left(t\right)\ , \tag{5.2.1}$$

$$\frac{dN}{dt} = \gamma_N\left(\mu - N - N\left|E\right|^2\right). \tag{5.2.2}$$

The parameters in Eqs. 5.2.1-5.2.2 are: κ, the field decay rate, which we set at $300\,\mathrm{ns}^{-1}$; α, the linewidth enhancement factor, which we set at 3; $\Delta\omega$, the optical frequency detuning, which we set at $2\pi \times 0.49\,\mathrm{GHz}$; P_{inj}, the optical injection strength, which we set to $60\,\mathrm{ns}^{-2}$; D, the noise level, which we varied; γ_N, the carrier decay rate, which we set at $1\,\mathrm{ns}^{-1}$, μ the pump current parameter, which we varied. $\xi(t)$ is a complex uncorrelated Gaussian noise of zero mean and unity variance that represents spontaneous emission: $\xi(t) = \xi_r(t) + i\xi_i(t)$ with $\langle\xi_r(t)\xi_r(t')\rangle = \delta(t-t')$, $\langle\xi_i(t)\xi_i(t')\rangle = \delta(t-t')$ and $\langle\xi_r(t)\xi_i(t')\rangle = 0$.

To simulate the evolution of Eqs. 5.2.1 and 5.2.2, we used the Runge-Kutta method of order 2 with a time step of 10^{-3} ns, as described in Ref. [191], which takes into account the stochastic evolution with white noise. In this work, we will analyze the chaotic pulses that appear at the

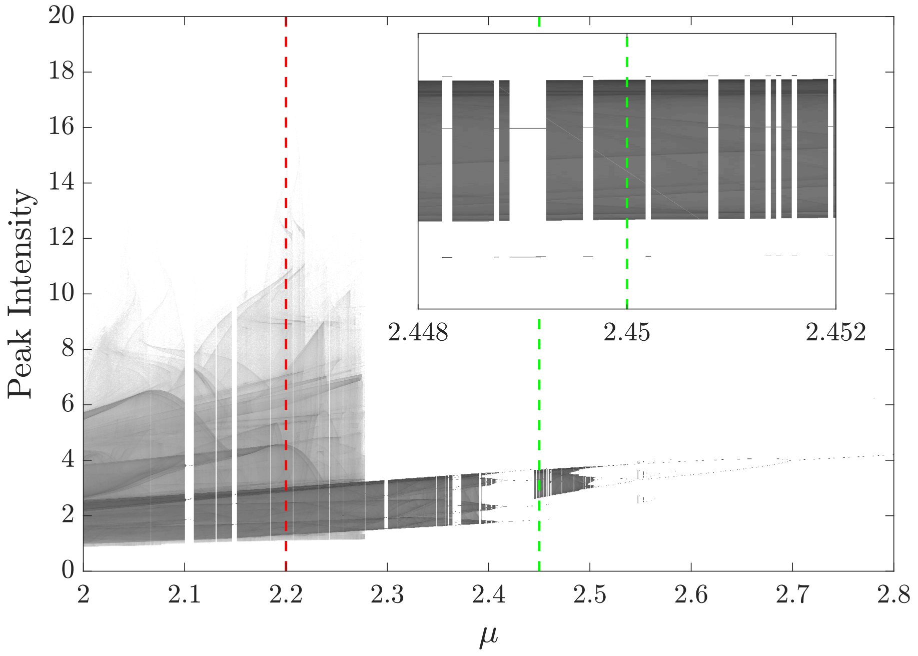


Figure 5.2.1.: Deterministic bifurcation diagram of the output intensity of the injected laser (D=0) when varying the pump current parameter μ. For the subsequent analysis, we choose two currents that lead to very different dynamical behaviors: (i) $\mu = 2.2$, where the system presents extreme events, and (ii) $\mu = 2.45$ where the systems displays a bounded chaotic behaviour. Example time series for these parameters, including noise in the simulations ($D = 10^{-4}\,\mathrm{ns}^{-1}$) are shown in Fig. 5.2.2.

output intensity of the laser defined as $P = |E|^2$.

Figure 5.2.1 displays how the intensity deterministic dynamics ($D = 0$) depends on the pump current parameter μ. For small μ (not shown), the laser emits a constant intensity, but as μ increases a Hopf bifurcation and a series of period-doubling bifurcations occur, resulting in chaotic emission. Around $\mu = 2.2$, the intensity shows extreme pulses as shown in Fig. 5.2.1 and the time series in Fig. 5.2.2(a and b). In contrast, at around $\mu = 2.45$, the amplitude of the pulses in this chaotic regime is tightly bounded as it can be seen in Fig. 5.2.1 and in the time series in Fig. 5.2.2(c and d).

The autocorrelation functions of the peak intensity values (i.e., the autocorrelation of the series y_i built with the amplitude of each intensity peak) for both values of μ and both values of D are shown in Fig. 5.2.2. For $\mu = 2.2$, the autocorrelation of the peak series decays to zero after a few peaks, both for $D = 0$ and $D = 10^{-4}\,\mathrm{ns}^{-1}$. It can be seen that for $\mu = 2.45$, the autocorrelation of the peak series does not decay to zero and shows non-negligible values of the autocorrelation even after 8 peaks. The values of the autocorrelation as a function of the time lag (number of peaks) are larger for the series with noise. We show in Fig. 5.2.2(d) that the evolution of the laser intensity with noise alternates regions of more regular behavior with regions of chaotic dynamics, which is not seen in the time series without noise in Fig. 5.2.2(c). This is due to the fact that $\mu = 2.45$ lies in a small chaotic island near regular regimes (see Fig. 5.2.1) and we find noise-induced jumps between different dynamical regimes. We anticipate that the faster decay of the autocorrelation function, together with the presence of extreme pulses, in the time series for $\mu = 2.2$ will result on

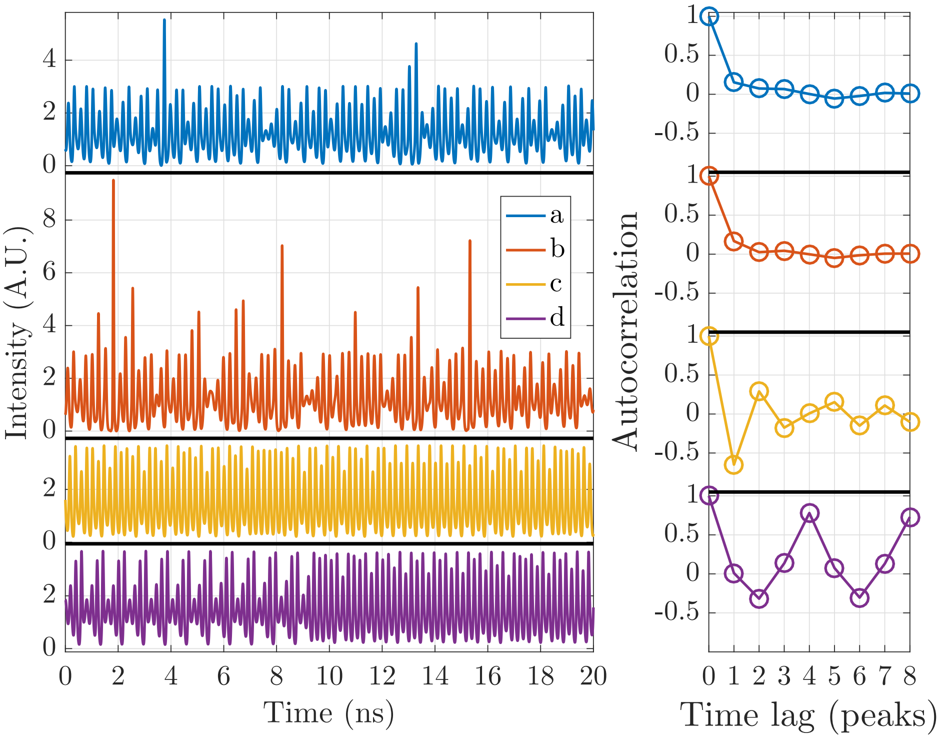


Figure 5.2.2.: (Left) Intensity time-series of the laser with optical injection and (right) autocorrelation function of the extracted peak series for $\mu = 2.2$ (a and b) and $\mu = 2.45$ (c and d) and noise level of $D = 0$ (a and c) and $D = 10^{-4}\,\mathrm{ns}^{-1}$ (b and d).

larger prediction errors than in the time series for $\mu = 2.45$.

5.3. Forecast methods

All machine learning methods used here tackle the problem of function approximation. We use them to forecast the amplitude of the upcoming intensity peaks by assuming that there is an objective function (that we try to infer) that takes as inputs a certain number of consecutive peak amplitudes and returns as output the amplitude of the next peak.

Except for the method of reservoir computing (that has an internal state with memory of the history of the inputs), all other methods are memoryless (i.e., they have no internal state of the history of the inputs), and explicit input and outputs of the objective function have to be provided in the training phase, providing information of the history with the previous intensity peaks amplitude. Let y_i be the i-th intensity peak amplitude, our objective function is

$$f\left(y_{i-n}, ..., y_{i-1}\right) = y_i, \tag{5.3.1}$$

where n is the number of input intensity peak amplitudes that the machine learning algorithm is fed with. For the forecast of the peak amplitudes, we found that keeping $n = 3$ yielded the minimum prediction error and further increasing n produced no accuracy enhancement. This choice will be justified in more detail in the results section (see Fig. 5.4.5).

For simplicity we call $\mathbf{x}_i = \left(y_{i-n}, ..., y_{i-1}\right)$ and thus, we can rewrite Eq. 5.3.1 as:

$$f\left(\mathbf{x}_i\right) = y_i. \tag{5.3.2}$$

For testing the methods, we use a different realization of the same simulations (not used in the training phase), and with this new data we evaluated the learned function,

$$\tilde{f}\left(\mathbf{x}_i\right) = \tilde{y}_i. \tag{5.3.3}$$

Several statistical measures have been used in the literature to quantify the performance of time series prediction algorithms such as the correlation coefficient (CC) [192], the mean squared error (MSE) [185], the normalized mean squared error (NMSE) [186], the root mean squared error (RMSE) [185], the normalized root-mean-square error (NRMSE) [187], the mean absolute relative error (MARE), etc. Here we use the MARE [192] defined as:

$$\mathrm{MARE} = \frac{1}{N}\sum_{i=1}^{i=N}\frac{|\tilde{y}_i - y_i|}{y_i}. \tag{5.3.4}$$

In the following subsections we describe the different algorithms used.

5.3.1. Statistical methods

5.3.1.1. k-Nearest Neighbors

The k-Nearest Neighbors (KNN) is a popular method used for supervised learning [193]. It works by finding, in the training set, the k most similar points to a test point. Then, the prediction of the test point is obtained by averaging the response of such k points (in the training set). Thus,

$$\tilde{y} = \frac{1}{k}\sum_{j\in\mathscr{N}} y_j, \tag{5.3.5}$$

where $\mathscr{N}$ (the neighborhood of the test point $\mathbf{x}_i$) is the set of indexes of the k points in the training set that are closest to the test point.

5.3.1.2. Support Vector Machine

Support Vector Machine [194–196] (SVM), is another popular method used for supervised learning, which is based on the inner product of points in the set to approximate the response function [197]. Nonlinearities can be introduced straight-forwardly by modifying the inner product function. For linear SVM the inner product of two points ($\mathbf{x}_i$ and $\mathbf{x}_j$) is calculated as

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^t \mathbf{x}_j, \tag{5.3.6}$$

while nonlinearity can be introduced by using a Gaussian kernel to calculate the inner product,

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right). \tag{5.3.7}$$

The objective function, $\tilde{f}\left(\mathbf{x}_i\right) = \tilde{y}_i$, is written as a linear combination of the inner products with the support vectors

$$\tilde{f}\left(\mathbf{x}_i\right) = \sum_j \beta_j \left\langle \mathbf{x}_j, \mathbf{x}_i \right\rangle + b. \tag{5.3.8}$$

The coefficients β_j and b are obtained by solving a convex optimization problem [195].

The linear SVM has the advantage of being parameter-free. In contrast, for using the Gaussian kernel the scale factor σ has to be defined. To set the value of σ we used the automatic heuristic implemented in the *Statistics and Machine Learning Toolbox* of MatLab (the *fitrsvm* function).

5.3.2. Artificial neural networks

5.3.2.1. Feed-forward neural networks

Feed-forward neural networks, usually simply referred as neural networks, use a set of units, called perceptrons, that, when used in a large network, their output can approximate a great variety of functions depending on the weights of the connections among the units.

Perceptrons perform two tasks, they compute a weighted sum of all their inputs (and a constant bias input), and they perform a nonlinear function, called activation function, to the result. The output of the activation function is the output of the perceptron. Most commonly, the activation functions used are sigmoids, in this work we use the tanh function in all but the last (output) layer, in which we don't use a nonlinearity to avoid bounding the final output to the codomain of the nonlinearity.

A feed-forward neural network, is a network of such perceptrons wherein they are ordered in layers, as shown in Fig. 5.3.1(a) for a single hidden layer. The perceptrons of the first layer have their inputs set to be the inputs of the whole network. For the rest of the layers, the inputs are defined as the outputs of the perceptrons in the previous layer.

The parameters of these networks are the weights of each perceptron. These parameters can be set using a gradient descend algorithm, in feed-forward neural networks an efficient algorithm to perform gradient descend, called back-propagation [198], may be used.

We used a shallow neural network (shallow NN), consisting of a single hidden layer of 30 perceptrons and a deep neural network (deep NN) consisting of 5 hidden layers of 10, 20, 50, 25, and 10 perceptrons (ordered from the input layer to the output layer), respectively.

5.3.2.2. Reservoir computing

Reservoir computing (RC) is a computational paradigm that can be viewed as a particular type of artificial neural networks with a single hidden layer and recurrent connections [199]. A ring topology in the hidden layer (or reservoir), as the one shown in Fig. 5.3.1(b), is a simple way to create recurrent connections. Such a ring topology yields a performance comparable to more complex network topologies in the reservoir [200]. Being a recurrent neural network, the reservoir computing technique is suitable to process sequential information. In reservoir computing, the connection weights from the input layer to the hidden layer as well as the connection weights within the reservoir are drawn from a Gaussian distribution and left untrained. The connection weights from the reservoir to the output layer are trained in a supervised learning procedure, which translates to a linear problem that can be solved via a simple linear regression [201].

The nodes in the reservoir layer perform a nonlinear transformation of the input data. Here we use a sine squared nonlinearity, which can be implemented in photonic hardware [202, 203], but other types of nonlinearity are also possible. Finally, the output node performs a weighted sum of the reservoir outputs.

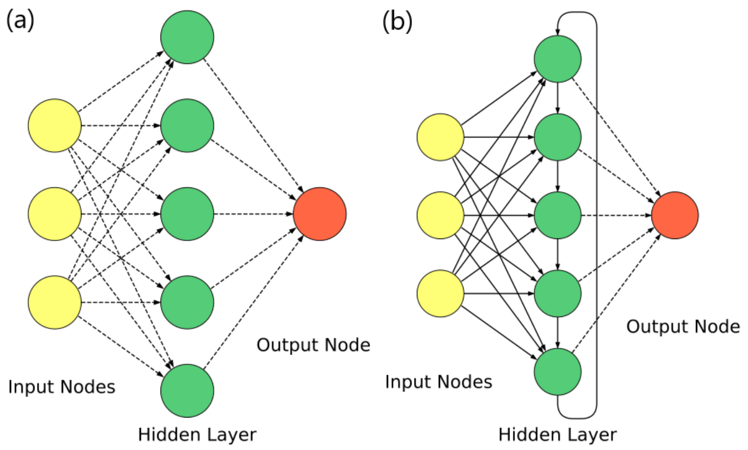


Figure 5.3.1.: Graphical representation of a feed-forward neural network with a single hidden layer (a) and a reservoir computer (b). The dashed lines represent the connections that can be adjusted via a learning procedure while the solid lines account for the connections that are randomly weighted and left untrained.

The RC method can be described by the following equations for the states of the nodes in the hidden layer (z^j) and the prediction of the output node ($\tilde{y}$):

$$z_i^j = F(\gamma w_j^I y_{i-1} + \beta z_{i-1}^{j-1}), \tag{5.3.9}$$

$$\tilde{y}_i = \sum_{j=1}^{D} w_j^O z_i^j; \tag{5.3.10}$$

where i refers to the peaks in the laser time series, j is the index of the node in the hidden layer, w^I are the set of input weights drawn from a random Gaussian distribution, γ and β are the input and feedback scaling, respectively, and $F(u) = \sin^2(u + \phi)$ is the nonlinear activation function. In Eq. 5.3.10, D is the number of hidden nodes and w^O stands for the trained output weights. In order to create a recurrent ring connectivity in the hidden layer (also known as reservoir), we connect node z^j ($j = \{2...D\}$) with its neighbor z^{j-1} and we close the ring by connecting node z^1 with z^D as shown in Fig. 5.3.1(b). Here, we have set the hyper-parameter values as $\gamma = 4.5$, $\beta = 0.25$, $\phi = 0.6\pi$ and $D = 6000$, which minimize the prediction error. We have verified that a tanh activation function yields quantitatively similar results once the hyper-parameters γ and β are optimized.

We note that the heuristic for the RC practitioners is to assume a random interconnection topology in the reservoir, which usually yields good results. However, regular network topologies also yield optimal results as long as the hyper-parameters are optimized [204, 205], as it has been the case here. For the RC method, we only feed a single amplitude value to predict the amplitude of the next pulse. Feeding the RC method with the value of several previous peaks would mean that, in practice, the reservoir computer would not need to use its own internal memory. The motivation to employ a different number of input peaks for the reservoir computer lies on the observation that it can reach a prediction error comparable to the other methods without using explicit memory of the preceding peaks.

5.4. Results and discussion

We now proceed to evaluate the performance of the different forecast methods on the prediction of the amplitude of chaotic laser pulses. The goal of our work is to predict the amplitude of the upcoming laser pulse given the recent history of the dynamics. To that end, we generate long time series of a laser subject to optical injection following the model described in Eqs. (5.2.1) and (5.2.2) for the two chaotic regimes shown in Fig. 5.2.2.

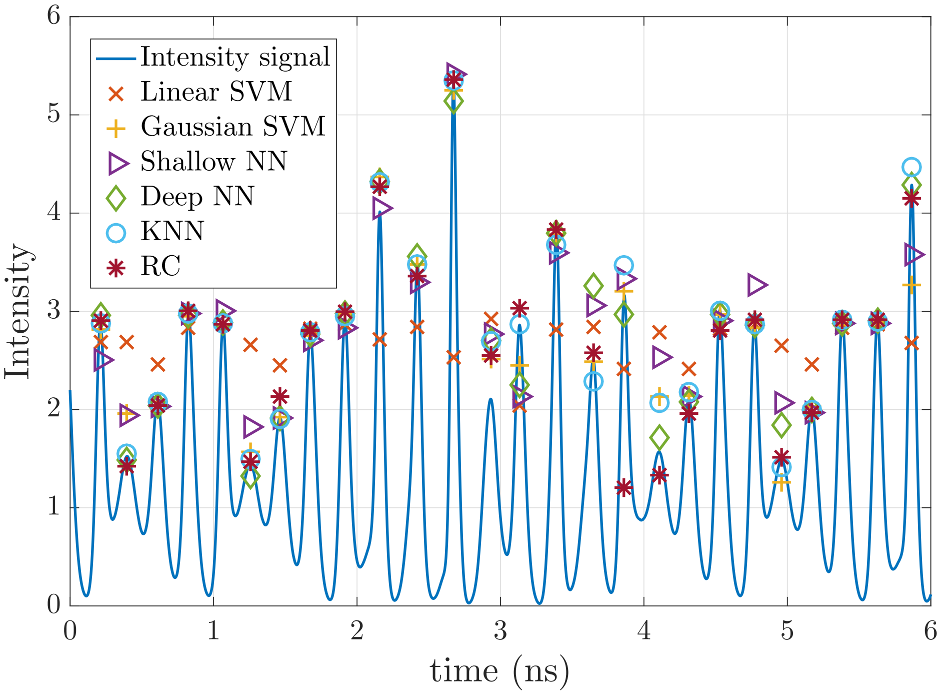


Figure 5.4.1.: Simulated intensity time series together with the peak amplitudes predicted by the different methods. The parameters are $D = 10^{-4}\,\mathrm{ns}^{-1}$ and $\mu = 2.2$. All methods were trained using 15000 ns of simulation, which contain 65534 peak intensity values.

By looking at Fig. 5.2.2, the presence of extreme events in the time series of the laser when the current is $\mu = 2.2$ becomes apparent. We anticipate that the existence of such extreme events poses a challenge for the prediction of the chaotic laser pulses' amplitude. Figure 5.4.1 shows a segment of the time series of the laser for the parameters $\mu = 2.2$ and $D = 10^{-4}\,\mathrm{ns}^{-1}$ together with the prediction of the pulses amplitude for all the methods considered in this work. From this first qualitative evaluation of the forecast methods, we can observe how the linear SVM method is outperformed by the other methods. In turn, the methods Deep NN, KNN and RC tend to yield a similar, accurate, prediction of the amplitude of the chaotic pulses.

A further visualization of the goodness of the different methods is provided by the scatter plots in Fig. 5.4.2. These scatter plots represent the predicted peak intensities versus the real ones. The methods with a better prediction accuracy need to align to a diagonal line in this representation. For this chaotic regime of the laser dynamics with the presence of extreme events, the Deep NN, KNN and RC methods are well aligned to the diagonal lines as shown in Figs. 5.4.2(d)-(f). In contrast, the Shallow NN and Gaussian SVM methods tend to underestimate the amplitude of medium to large pulses as it can be seen in Fig. 5.4.2(b)-(c). As shown in Fig. 5.4.2(a), the linear SVM method fails to capture the complexity of the dynamics.

When doing data-driven forecasting, it is necessary to evaluate the number of training points

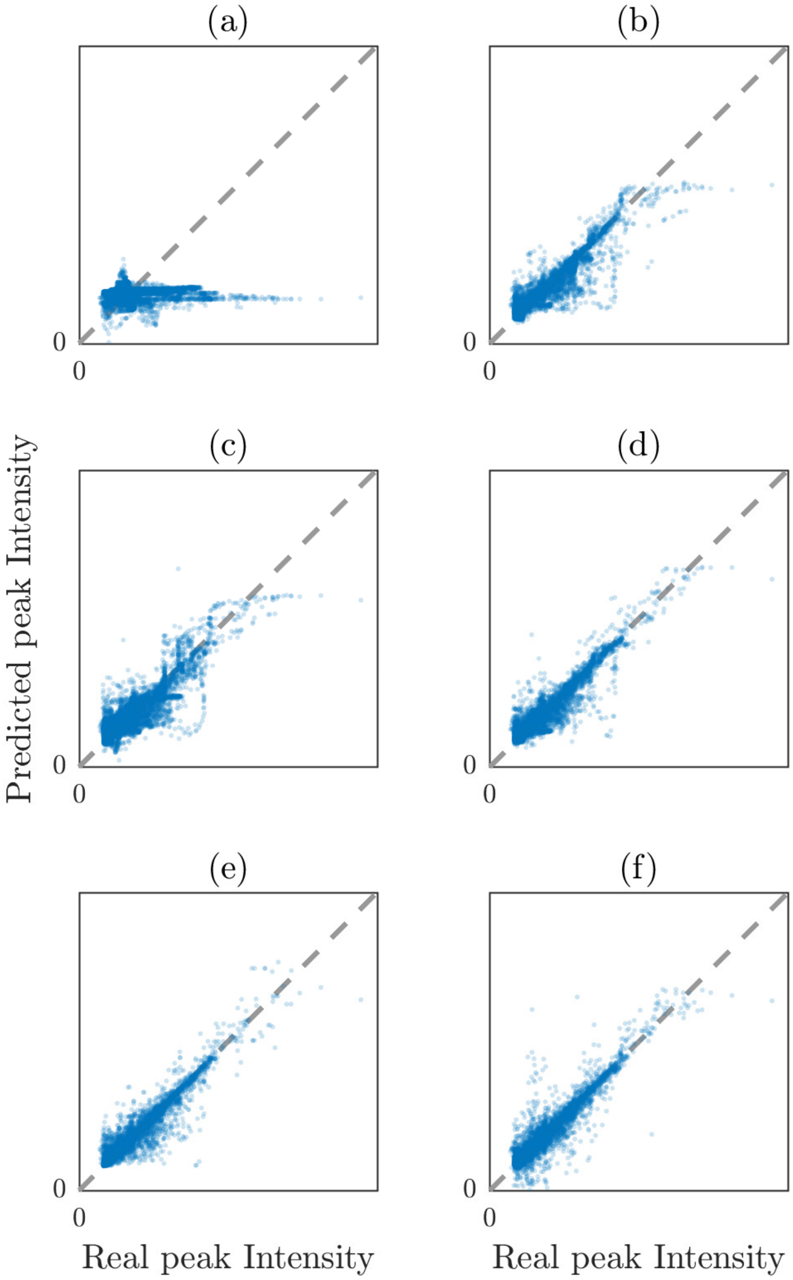


Figure 5.4.2.: Scatter plots displaying the simulated peak intensity *vs.* the predicted peak intensity for the methods (a) Linear SVM, (b) Gaussian SVM, (c) Shallow Neural Network, (d) Deep Neural Network, (e) k-Nearest Neighbors, (f) Reservoir Computing. The parameters are $D = 10^{-4}\,\text{ns}^{-1}$ and $\mu = 2.2$. All methods were trained using 15000 ns of simulation, containing 65534 peak intensity values.

needed to have accurate results. In Figs. 5.4.3 and 5.4.4 we show how the accuracy (as measured by the mean absolute relative error, Eq. 5.3.4) depends on the number of points used to train the algorithms, when there are extreme pulses (Fig. 5.4.3) and when there are no extreme pulses (Fig. 5.4.4).

First, we compare the forecast results for the noise-free numerical simulations at currents $\mu = 2.2$

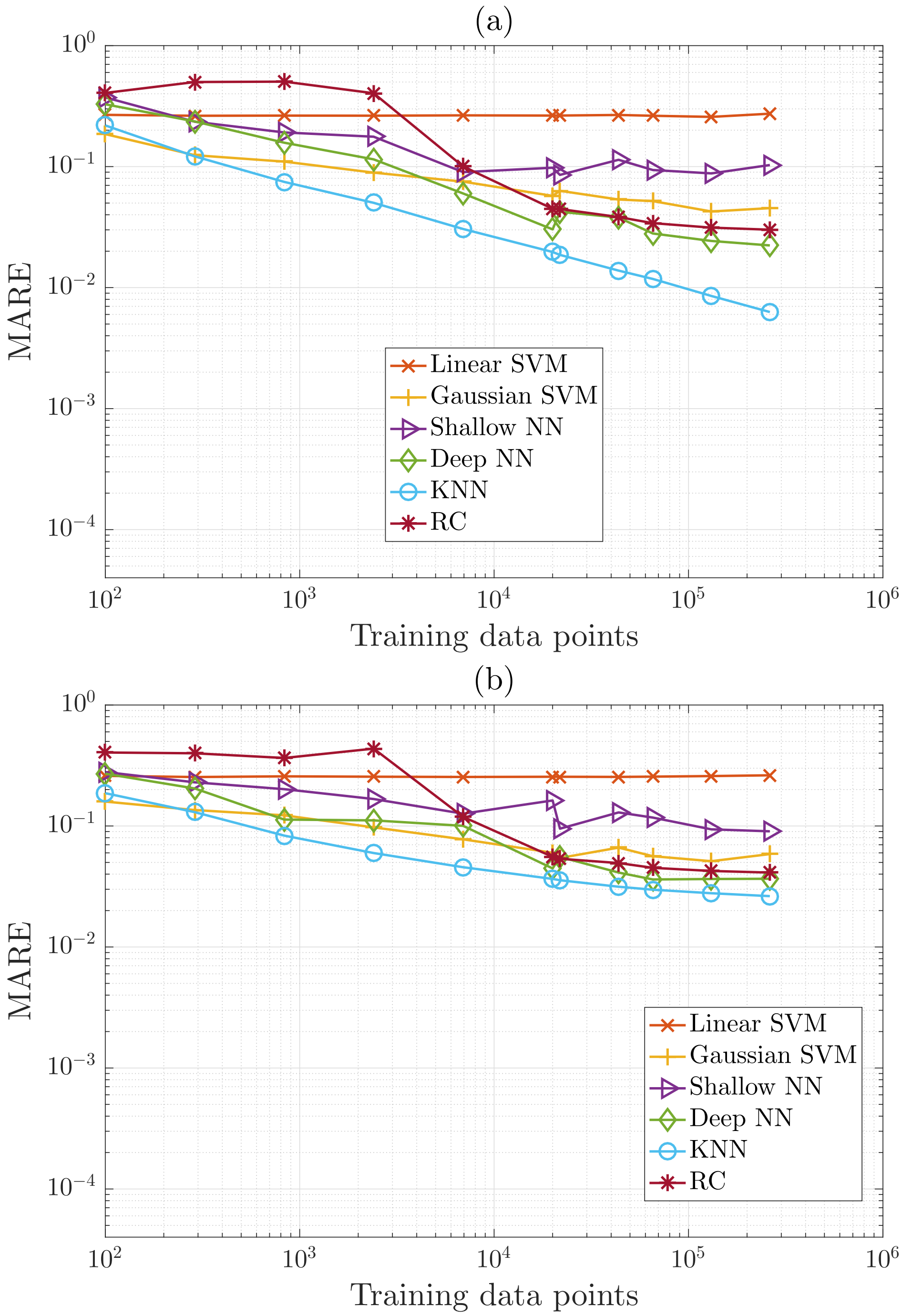


Figure 5.4.3.: Mean absolute relative error as a function of number of training points. We show the error of the peak amplitude prediction as a function of the number of training points for noise levels of (a) $D = 0$ and (b) $D = 10^{-4}\,\mathrm{ns}^{-1}$, at $\mu = 2.2$.

and $\mu = 2.45$, which are shown in Figs. 5.4.3(a) and 5.4.4(a). The MARE of the forecast for $\mu = 2.2$ is at least two orders of magnitude worse than the forecast for $\mu = 2.45$. This is due to the added complexity of the extreme events at $\mu = 2.2$, deteriorating the performance of all the forecasting methods. We find that the KNN, Deep NN, and RC methods, in this order, yield the most accurate predictions for $\mu = 2.2$. These methods, together with the Shallow NN, yield the lowest MARE for $\mu = 2.45$. In both cases, the performance of the RC method becomes more accurate when the number of training data points is larger than the number of nodes in the reservoir ($D = 6000$).

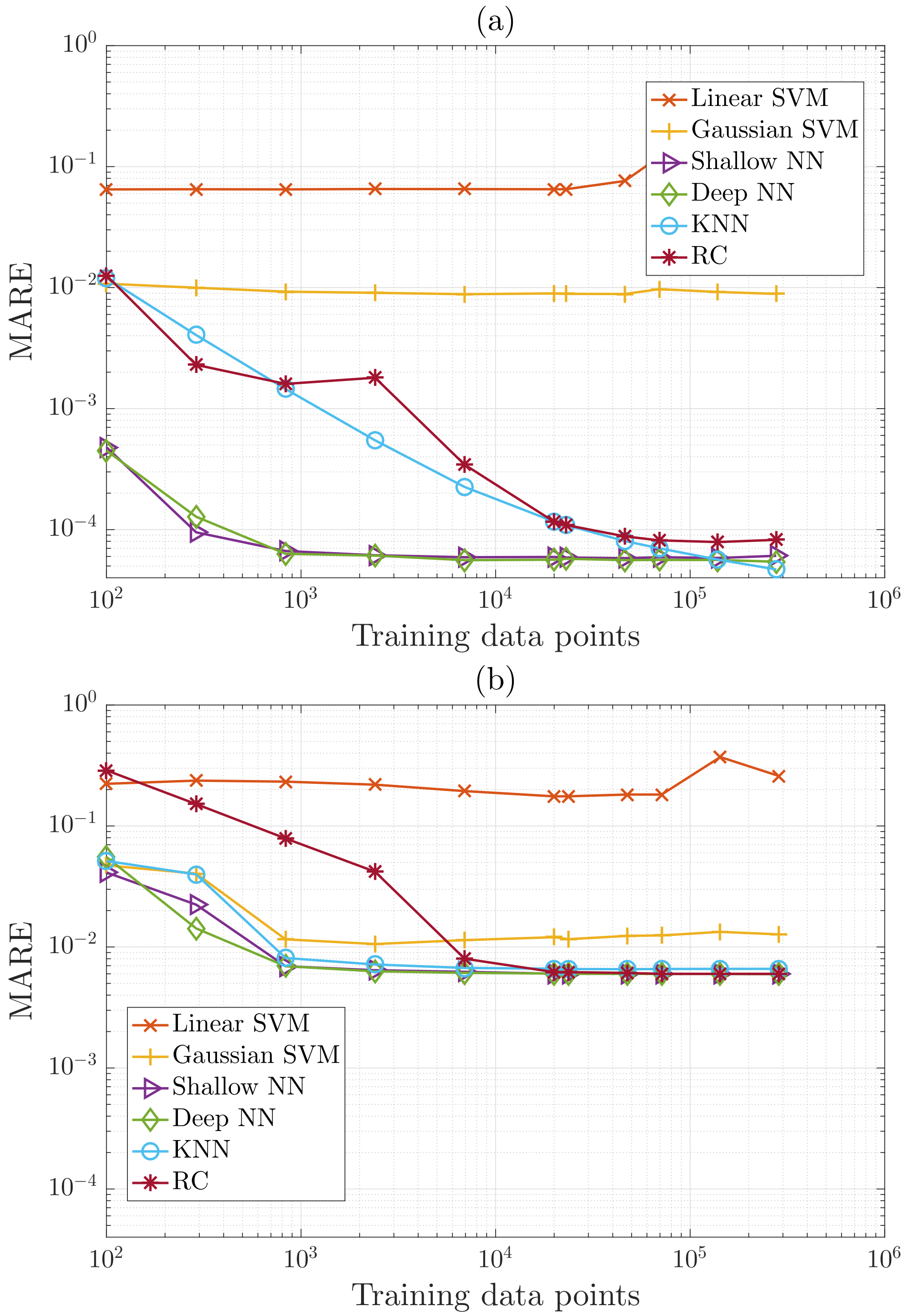


Figure 5.4.4.: Mean absolute relative error as a function of number of training points. We show the error of the peak amplitude prediction as a function of the number of training points for noise levels of (a) $D = 0$ and (b) $D = 10^{-4}\,\text{ns}^{-1}$, at $\mu = 2.45$.

Overall, the prediction of the amplitude of the upcoming chaotic pulse for $\mu = 2.45$ requires less training points than for $\mu = 2.2$. These results suggest that the forecast of the dynamics with extreme pulses is intrinsically harder to predict. It could also be that the low frequency of the extreme pulses makes them more difficult to predict because they appear less frequently in the training set. However, they also appear less frequently in the testing set and thus have less weight in the overall error.

Second, we analyze the influence of the stochastic contribution in Eq. 5.2.1 on the forecast of the

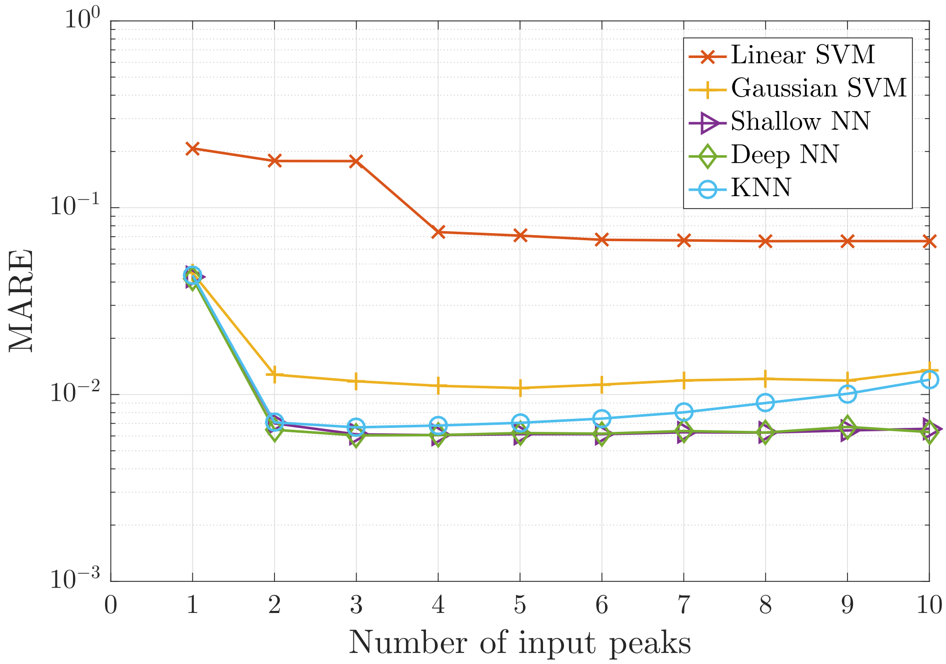


Figure 5.4.5.: Mean absolute relative error as a function of the number of preceding pulses fed as the input to each algorithm. For this example, all algorithms were trained using 10000 data points at $\mu = 2.45$ and $D = 10^{-4}\,\text{ns}^{-1}$. For both KNN and Deep NN the minimum error occurs when using 3 input peaks, being the latter the absolute minimum in this plot considering all other methods.

pulses' amplitude. We show in Figs. 5.4.3(b) and 5.4.4(b) that the presence of noise triggers an early plateau that bounds the MARE, deteriorating the performance of all the methods. The stochastic contribution to the dynamics has a stronger influence on the forecast for the chaotic dynamics generated at $\mu = 2.45$, with an increase of two orders of magnitude in the MARE as shown in Figs. 5.4.4 (a) and (b). When noisy dynamics is considered, the MARE for $\mu = 2.45$ and $\mu = 2.2$ are less than an order of magnitude apart (see Figs. 5.4.3(b) and 5.4.4(b)) in contrast to the noise-free counterparts for which the difference in MARE between $\mu = 2.45$ and $\mu = 2.2$ is more apparent (see Figs. 5.4.3(a) and 5.4.4(a)). The deterioration of the prediction accuracy in the presence of observational noise has also been reported, e.g., in Ref. [186], where the NMSE decreased 5 to 6 orders of magnitude.

An important parameter for all but the reservoir computing approach is the number of input intensity peak amplitudes (n) wherewith the machine learning algorithm is fed. This parameter sets the amount of history that the algorithm is able to "see". In Fig. 5.4.5, we show how the performance changes when changing n in the case of the chaotic dynamics with $\mu = 2.45$ and $D = 10^{-4}\,\text{ns}^{-1}$. We used 10000 training data points to be well inside the plateau of performance seen in Fig. 5.4.4(b). The results shown in this figure justifies our choice of using $n = 3$, which yields a minimum MARE for most of the forecast methods. For the RC method, we set $n = 1$ as it is the only method that possesses an internal memory.

Another important issue to consider when implementing these data-driven methods is the computer power that is required to train and test each forecast method. Although different methods scale differently with the amount of data in the training set, some general rules of thumb apply. In the KNN method, while there is no specific training time, the time for evaluating each test point, however, grows linearly with the amount of points in the dataset. The KNN method is, in this sense, ideal for real-time data as it can take into account new data into the dataset without any extra computational overhead. The computing power for training and testing the SVM methods depends greatly on the amount of support vectors that are needed, and on the kernel that is used. We find that the linear SVM method takes a greater time to train and a comparable time to test with respect to the Gaussian kernel method. This is due to the fact that the linear SVM method fails to capture the complexity of the data and, thus, a great amount of support vectors are needed. The feed-forward neural networks and the RC method are (in respect to train and test) opposite to the KNN method, they take a great amount of time to train but are computationally cheap to evaluate test points. The training time of the neural networks-based models depends on the length of the training data and on the amount of internal model parameters they have. In our examples, the deep NN takes about 20 times the time of the shallow NN to train. The RC method, on the other hand, has a simpler training mechanism, which takes approximately 5 times the time of the shallow NN to train. All neural networks-based models take a comparable (low) time in the testing stage.

We end up with a comparison of the performances obtained here with the literature. The reported MARE values that have been obtained strongly vary with the algorithm used and the characteristics of the datasets analyzed. For example, machine learning techniques with delay embedding in real data give MARE values of the order of 0.15 for river flow prediction [192], or as low as 0.025 for electricity consumption [184]. With time series simulated from the chaotic Ikeda map, a MARE value as low as 5.8×10^{-5} was reported in Ref. [206]. In general, the prediction of noisy (possibly chaotic) real-world dynamics yields larger errors than the prediction of synthetic numerical data without noise. A more precise direct comparison of previously published results is, however, not currently possible since we do not predict the future trajectory of the dynamics but the amplitude for the next pulse.

5.5. Conclusions

We have used the chaotic dynamics of the intensity of an optically injected laser to test the performance of several machine learning algorithms for forecasting the amplitude of the next intensity pulse. This laser system is described by a simple model that, with a small change of parameters, produces time series which have extreme events in the form of high peak intensities, resembling the dynamics of much more complex systems. In spite of the fact that the autocorrelation function of the sequence of pulse amplitudes decays rapidly, good prediction accuracy was achieved with some of the proposed methods, namely the KNN, Deep NN and RC methods. We have verified that the MARE for the most accurate methods (DNN, KNN and RC) remains approximately constant even for the prediction of extreme pulses that have a probability of appearance as low as 1/1000.

Our work suggests that similar methods may be used in the forecast of more complex systems, although further testing is of course necessary to asses how well they would perform in high-dimensional chaotic dynamical systems. While with simple dynamics, we only needed around 1000 data points to achieve maximum performance with some methods (shallow and deep NN); when forecasting more complex dynamics, some of the methods (KNN and deep NN) will continue improving their performance if longer datasets are available for training (longer than 10^5 data points).

We have also compared the performance of different variations of the same machine learning algorithm (compare linear to Gaussian SVM and shallow to deep neural networks in Figs. 5.4.1-5.4.4), especially relevant when considering big training datasets. We do not exclude that even more complex methods (e.g., a neural network with additional hidden layers) might outperform the presented algorithms. However, the presented algorithms already serve the purpose of showing the dependence of the forecast error on the complexity of the dynamics and on the inclusion of stochastic contributions.

Chapter 6.

Discussion of the results

The work presented in this thesis was aimed at developing reliable algorithms to analyze complex datasets with special focus in ophthalmic images.

In chapter 2 (Ref. [7]) we developed an unsupervised procedure for ordering anterior segment OCT images according to the iris-corneal angle, which is the main risk factor of angle-closure glaucoma. Such ordering can allow to classify the images in meaningful groups according to relevant features. We tested the algorithm with two sets of images, one classified by two expert ophthalmologists, and another with a larger set of annotated images. We obtained high correlations (0.80∼0.90) between the features extracted by our unsupervised ordering algorithm and the features from the manual annotation or classification of images.

The abstract features generated by the algorithm provide novel tools for assessing OCT images of the anterior chamber. They can be used for direct classification of the images and, furthermore, they can be linked to established quantities used for characterizing diseased eyes (like chamber depth, iris-corneal angle) resulting in an automatic detection system. As the algorithm is fully unsupervised, it can be incorporated in OCT imaging systems to aid technicians and doctors in an early diagnosis.

The two main advantages of the algorithm, over previous works, demonstrated in our research are that it is fully unsupervised, and it does not rely on specific landmarks (other authors [60] propose first a segmentation to locate specific landmarks, such as the scleral spur, and use them to draw the conclusions); therefore, our algorithm can analyze images in which relevant landmarks are not visible or not easy to locate.

In chapter 3 (Ref. [8]) we developed two novel outlier finding techniques that can improve the above mentioned correlations. These techniques are also unsupervised, and thus, can be used before any other machine learning algorithm. Because they rely, only, on the distances between entries on a database, they can be used on any kind of data as long as a meaningful distance can be defined, which makes them ideal for very high-dimensional data such as OCT images. We tested these techniques not only with the anterior chamber OCT database, but with other real and synthetic databases, obtaining similar or even better performance than other outlier finding techniques previously proposed in the literature. In particular, we tested our techniques for detecting fraudulent transactions in a credit card transaction database, obtaining areas under the receiver operating characteristic in the test sets usually above 0.9 and average precision in the order of ∼0.1 (see Fig. 3.4.3). The main drawback of the proposed methods is that they scale badly with the number of entries in the database analyzed, as the distance matrix (of size N^2) needs to be computed and analyzed. Their performance varies with the dataset analyzed because the underlying assumption of what an outlier is, is different for the different methods. The percolation-based method assumes that the normal elements will be in one large cluster, with outliers being far from that cluster; the IsoMap-based method assumes that the normal elements lie on a manifold, and that outliers lie outside such manifold.

As stated before, the algorithms developed in chapter 3 (Ref. [8]), can improve the results of the OCT ordering algorithm, but they can also be used to assist any other algorithm that uses real-world databases possessing outlying entries or images with artifacts. Also, outliers found in medical images could be an indicator of a disease, although additional studies are needed in order to asses the proposed algorithms capabilities for such a task. Both proposed methods are completely unsupervised, one of them is parameter-free, meaning that it can be used blindly in any database, while the other has two integer number parameters that can be selected to optimize its performance using a training set.

In chapter 4 (Ref. [9]) we proposed new network features that, when calculated on vascular networks of human retinas, can distinguish between healthy eyes, eyes with glaucoma, and eyes with diabetic retinopathy. In particular, we showed how C-NDD (with $l = 1$ and $a = 2$) and CMWD (with $l = 0$ and $a = -2$) can distinguish the healthy groups from the glaucoma group (with p-values below 0.0066, see Table 4.2). While C-NDD (with $l = 1$ and $a = -2$) and WDD (with $l = 0$ and $a = 1$) can distinguish the healthy groups from the diabetic retinopathy group (with p-values below 0.011, see Table 4.1). The parameters l and a chosen to test the network analysis have all a clear physical interpretation. The set $l = 0$ and $a = 1$ is simply the width of the corresponding vessel. The set $l = 1$ and $a = 2$ is proportional to the volume of such vessel. And finally, the set $l = 1$ and $a = -2$ can be related to the flow resistance of the vessel.

In most cases, the algorithms performed better when using the HRF manual segmentation compared to the automatic segmentation. We think that this is due to the intrinsic better quality of the manual segmentation. It should also be noted that the analyzed network is a 2-dimensional projection of the real 3-dimensional retina network, this implies that there are some nodes in it which, in reality, correspond to crossovers of veins and arteries. This alters the extracted features in two ways, by generating spurious nodes, and by generating spurious shortest paths to the optical nerve. We have tested an algorithm that eliminates spurious nodes (for example, those that have 4 links), but we found that the modification did not improve the performance of the proposed measures, while it added complexity and more parameters to the algorithm.

To complement the work done in image analysis, in chapter 5 (Ref. [10]) we analyzed machine learning algorithms for time series prediction. We compared the performance of different machine learning methods to forecast a simulated chaotic laser signal. We changed a control parameter of the system that let us tune the complexity of the laser signal, and we studied how the stochastic contribution to the laser dynamics affects the quality of the forecast. In our best results, we obtained MARE (Mean Absolute Relative Error) of less than 10^{-4}, this was achieved for the low-complexity deterministic system, on the other hand, when considering high-complexity stochastic system, the MARE was around 0.03. In all cases we studied how much training data is needed to obtain accurate results, showing that the presence of noise triggers an early plateau in the plot of MARE *vs.* training data-points. While with more regular dynamics, we only needed around 1000 data points to achieve maximum performance with some methods (shallow and deep NN); when forecasting more irregular dynamics, some of the methods (KNN and deep NN) continue to improve their performance if longer datasets are given for training (longer than 10^5 data points), this is because more data provides a better resolution of the structure of the attractor.

Chapter 7.

Conclusions and future work

In our research we tackled the problem of developing new tools to aid decision-making in the clinics, by unsupervisedly analyzing ophthalmic images and drawing automatic conclusions. Our results can help the doctors provide faster and more accurate diagnosis to the patients, which can ultimately translate into better life quality. We have also contributed to the fields of machine learning and complex data analysis. We have developed techniques that can unsupervisedly find outliers, or anomalous data in a database, which can be useful not only to improve the performance of machine learning algorithms, but in general, to find entries on a database that are fundamentally different to the rest of the entries, which can mean that they posses some anomaly that might help finding rare diseases. We have also used machine learning techniques to forecast the evolution of a stochastic chaotic system.

The algorithms proposed in chapter 2 (Ref. [7]) can be installed to automatically run in OCT instruments, making it possible to flag patients with a risk of developing angle-closure glaucoma, even when taking the exam for other reason, such as evaluation for refractive surgery. It can also lead to a more efficient treatment for urgent patients prioritizing them in accordance to the outcome of the algorithm. A patent protecting this algorithms has already been granted (see appendix D.2.1) and future work should be aimed at developing a user-friendly interface that can allow this innovations to be available at the clinic.

In chapter 3 (Ref. [8]), we have proposed two outlier finding techniques. We showed how they outperform or perform comparable to previous techniques proposed in the literature. Because the information these techniques use is only the distance between elements of a database, their applicability is very broad. We have shown that they can enhance the performance of the algorithms proposed in chapter 2, but also, how they are useful in other tasks, such as fraud detection. Future work will be aimed at testing new applications for these techniques, automatic quality assessment for OCT imaging systems is a promising application, a patent concerning these outlier mining techniques has also been issued (see appendix D.2.2). More broadly, future work should asses which other machine learning algorithms may be enhanced by eliminating the outliers found by our algorithms in their training sets.

As OCT technology is still expensive, it is important to analyze cheaper options, such as fundus retinal images. We have shown that the network features proposed in chapter 4 (Ref. [9]), are useful for classifying healthy from unhealthy patient groups. For both studied diseases, the proposed network features are able to separate the healthy group and the unhealthy groups with good statistical significance. We have also compared our results with those obtained from fractal geometry analysis, and we have shown that using the fractal dimensions over two versions of the same image (raw and skeletonized segmentations), improves the separation between the groups, in comparison to using only one (see Fig. 4.4.1). The most statistically significant results were obtained using high resolution images (the HRF database).

We also expect that other diseases unrelated to the eyes but to the circulatory system can become apparent using these techniques, as the vessel network in the retina is the easiest to image. Future work, with bigger datasets and a wider variety of diseases should asses the potential of the proposed features in a more general scenario.

Complementing the work done with image analysis, in chapter 5 (Ref. [10]) we focused on the prediction of a chaotic time series, specifically we used a laser system model that produces time series which have extreme events in the form of high peak intensities, resembling the dynamics of much more complex systems. In spite of the fact that the autocorrelation function of the sequence of pulse amplitudes decays rapidly, our results indicate that an accurate prediction of the amplitude of upcoming chaotic pulses is possible using machine learning techniques, although the presence of extreme events in the time series and the consideration of stochastic contributions in the laser model bound the accuracy that can be achieved.

The tools used to predict the behavior of a laser with optical injection in chapter 5 (Ref. [10]) could be useful for analyzing other time-series such as ECG [207]. This was, in fact, our first motivation for our study: to first test the performance in synthetic data (which we can easily control) before applying the techniques to real (medical) data. Thus, exploring the boundaries and constraints of such machine learning techniques can help us better understand their limitations. Our work suggests that similar methods may be used in the forecast of more complex systems, although further testing is of course necessary to asses how well they would perform in high-dimensional dynamical systems.

Bibliography

[1] Parlitz U., Berg S., Luther S. et al. Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics // Computers in Biology and Medicine. — 2012. — Vol. 42, no. 3. — P. 319 – 327.

[2] Long E., Lin H., Liu Z. et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts // Nature biomedical engineering. — 2017. — Vol. 1, no. 2. — P. 0024.

[3] Wang Y., Zhang Y., Yao Z. et al. Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images // Biomed. Opt. Express. — 2016. — Vol. 7, no. 12. — P. 4928–4940.

[4] Abu-Mostafa Y. S., Magdon-Ismail M., Lin H.-T. Learning from data. — AMLBook New York, NY, USA:, 2012. — Vol. 4. — ISBN: 1600490069.

[5] Friedman J., Hastie T., Tibshirani R. The elements of statistical learning. — Springer series in statistics New York, 2001. — Vol. 1. — ISBN: 0387952845.

[6] Brunton S. L., Kutz J. N. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. — Cambridge University Press, 2019. — ISBN: 1108422098.

[7] Amil P., González L., Arrondo E. et al. Unsupervised feature extraction of anterior chamber OCT images for ordering and classification // Scientific reports. — 2019. — Vol. 9, no. 1. — P. 1157.

[8] Amil P., Almeira N., Masoller C. Outlier mining methods based on graph structure analysis // Frontiers in Physics. — 2019. — Vol. 7. — P. 194.

[9] Amil P., Reyes-Manzano C. F., Guzman-Vargas L. et al. Network-based features for retinal fundus vessel structure analysis // PloS one. — 2019. — Vol. 14, no. 7. — P. 1–15.

[10] Amil P., Soriano M. C., Masoller C. Machine learning algorithms for predicting the amplitude of chaotic laser pulses // Chaos: An Interdisciplinary Journal of Nonlinear Science. — 2019. — Vol. 29, no. 11. — P. 113111.

[11] Tenenbaum J. B., De Silva V., Langford J. C. A global geometric framework for nonlinear dimensionality reduction // Science. — 2000. — Vol. 290, no. 5500. — P. 2319–2323.

[12] Dijkstra E. W. A note on two problems in connexion with graphs // Numerische mathematik. — 1959. — Vol. 1, no. 1. — P. 269–271.

[13] Floyd R. W. Algorithm 97: Shortest path // Commun. ACM. — 1962. — Vol. 5, no. 6. — P. 345–.

[14] Borg I., Groenen P. J. F. Modern Multidimensional Scaling. — 2 edition. — Springer-Verlag New York, 2005. — ISBN: 1441920463.

[15] Maaten L. v. d., Hinton G. Visualizing data using t-SNE // Journal of Machine Learning Research. — 2008. — Vol. 9, no. Nov. — P. 2579–2605.

[16] Saito T., Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets // PloS one. — 2015. — Vol. 10, no. 3. — P. e0118432.

[17] Nuzzo R. Scientific method: statistical errors // Nature News. — 2014. — Vol. 506, no. 7487. — P. 150–152.

[18] Krzywinski M., Altman N. Significance, P values and t-tests // Nature Methods. — 2013. — Vol. 10, no. 11. — P. 1041–1042.

[19] Wasserstein R. L., Lazar N. A. The ASA statement on p-values: Context, process, and purpose // The American Statistician. — 2016. — Vol. 70, no. 2. — P. 129–133.

[20] Newman M. E. J. The structure and function of complex networks // SIAM Review. — 2003. — Vol. 45, no. 2. — P. 167–256.

[21] Bennett T. J., Barry C. J. Ophthalmic imaging today: an ophthalmic photographer's viewpoint–a review // Clinical & Experimental Ophthalmology. — 2009. — Vol. 37, no. 1. — P. 2–13.

[22] Yannuzzi L. A., Ober M. D., Slakter J. S. et al. Ophthalmic fundus imaging: today and beyond // American Journal of Ophthalmology. — 2004. — Vol. 137, no. 3. — P. 511 – 524.

[23] Bernardes R., Serranho P., Lobo C. Digital ocular fundus imaging: A review // Ophthalmologica. — 2011. — Vol. 226, no. 4. — P. 161–181.

[24] Wojtkowski M., Srinivasan V., Fujimoto J. G. et al. Three-dimensional retinal imaging with high-speed ultrahigh-resolution optical coherence tomography // Ophthalmology. — 2005. — Vol. 112, no. 10. — P. 1734 – 1746.

[25] Aramendía A. R., Grulkowski I., Villar A. J. et al. Optimization of a SS-OCT with a focus tunable lens for enhanced visualization of ocular opacities // Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXIII / International Society for Optics and Photonics. — Vol. 10867. — 2019. — P. 108673E.

[26] Alterini T., Díaz-Doutón F., Burgos-Fernández F. J. et al. Fast visible and extended near-infrared multispectral fundus camera // Journal of biomedical optics. — 2019. — Vol. 24, no. 9. — P. 096007.

[27] Maamari R. N., Keenan J. D., Fletcher D. A., Margolis T. P. A mobile phone-based retinal camera for portable wide field imaging // British Journal of Ophthalmology. — 2014. — Vol. 98, no. 4. — P. 438–441.

[28] Budai A., Bock R., Maier A. et al. Robust vessel segmentation in fundus images // International Journal of Biomedical Imaging. — 2013. — Vol. 2013.

[29] Decencière E., Zhang X., Cazuguel G. et al. Feedback on a publicly distributed image database: the messidor database // Image Analysis & Stereology. — 2014. — Vol. 33, no. 3. — P. 231–234.

[30] Dal Pozzolo A., Caelen O., Johnson R. A., Bontempi G. Calibrating probability with undersampling for unbalanced classification // 2015 IEEE Symposium Series on Computational Intelligence / IEEE. — 2015. — P. 159–166.

[31] Dal Pozzolo A., Caelen O., Le Borgne Y.-A. et al. Learned lessons in credit card fraud detection from a practitioner perspective // Expert systems with applications. — 2014. — Vol. 41, no. 10. — P. 4915–4928.

[32] Dal Pozzolo A., Boracchi G., Caelen O. et al. Credit card fraud detection: a realistic modeling and a novel learning strategy // IEEE transactions on neural networks and learning systems. — 2018. — Vol. 29, no. 8. — P. 3784–3797.

[33] Dal Pozzolo A. Adaptive machine learning for credit card fraud detection. — 2015.

[34] Carcillo F., Dal Pozzolo A., Le Borgne Y.-A. et al. SCARFF: a scalable framework for streaming credit card fraud detection with spark // Information fusion. — 2018. — Vol. 41. — P. 182–194.

[35] Carcillo F., Le Borgne Y.-A., Caelen O., Bontempi G. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization // International Journal of Data Science and Analytics. — 2018. — Vol. 5, no. 4. — P. 285–300.

[36] Samaria F. S., Harter A. C. Parameterisation of a stochastic model for human face identification // Proceedings of 1994 IEEE Workshop on Applications of Computer Vision / IEEE. — 1994. — P. 138–142.

[37] Ju F., Sun Y., Gao J. et al. Image outlier detection and feature extraction via L1-Norm-Based 2D probabilistic PCA // IEEE Trans. Image Processing. — 2015. — Vol. 24, no. 12. — P. 4834–4846.

[38] Ohtsubo J. Semiconductor lasers: stability, instability and chaos. — Springer, 2017. — ISBN: 3319561375.

[39] Perrone S., Vilaseca R., Zamora-Munt J., Masoller C. Controlling the likelihood of rogue waves in an optically injected semiconductor laser via direct current modulation // Physical Review A. — 2014. — Vol. 89, no. 3. — P. 033804.

[40] Nitzberg M., Shiota T. Nonlinear image filtering with edge and corner enhancement // IEEE Transactions on Pattern Analysis & Machine Intelligence. — 1992. — no. 8. — P. 826–833.

[41] de Santos-Sierra D., Sendiña-Nadal I., Leyva I. et al. Graph-based unsupervised segmentation algorithm for cultured neuronal networks' structure characterization and modeling // Cytometry Part A. — 2015. — Vol. 87, no. 6. — P. 513–523.

[42] Zitová B., Flusser J. Image registration methods: a survey // Image and Vision Computing. — 2003. — Vol. 21, no. 11. — P. 977 – 1000.

[43] Gonzalez R. C., Woods R. E. et al. Digital image processing. — 3 edition. — Pearson, 2007. — ISBN: 9780131687288.

[44] Perona P., Malik J. Scale-space and edge detection using anisotropic diffusion // IEEE Trans. on Pattern Anal. and Machine Intell. — 1990. — Vol. 12, no. 7. — P. 629–639.

[45] Gerig G., Kubler O., Kikinis R., Jolesz F. A. Nonlinear anisotropic filtering of MRI data // IEEE Trans. on Medical Imaging. — 1992. — Vol. 11, no. 2. — P. 221–232.

[46] Sajda P. Machine learning for detection and diagnosis of disease // Annu. Rev. Biomed. Eng. — 2006. — Vol. 8. — P. 537–565.

[47] Bizios D., Heijl A., Hougaard J. L., Bengtsson B. Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT // Acta Ophthalmologica. — 2010. — Vol. 88, no. 1. — P. 44–52.

[48] Bowd C., Goldbaum M. H. Machine learning classifiers in glaucoma // Opt. and Vision Science. — 2008. — Vol. 85, no. 6. — P. 396–405.

[49] Bowd C., Hao J., Tavares I. M. et al. Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes // Investigative Ophthalmology & Visual Science. — 2008. — Vol. 49, no. 3. — P. 945–953.

[50] Burgansky-Eliash Z., Wollstein G., Chu T. et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study // Investigative Ophthalmology & Visual Science. — 2005. — Vol. 46, no. 11. — P. 4147–4152.

[51] Tham Y.-C., Li X., Wong T. Y. et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis // Ophthalmology. — 2014. — Vol. 121, no. 11. — P. 2081–2090.

[52] Mills R. P., Budenz D. L., Lee P. P. et al. Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease // Am. Journal of Ophthalmology. — 2006. — Vol. 141, no. 1. — P. 24–30.

[53] Nolan W. P., See J. L., Chew P. T. et al. Detection of primary angle closure using anterior segment optical coherence tomography in Asian eyes // Ophthalmology. — 2007. — Vol. 114, no. 1. — P. 33–39.

[54] Friedman D. S., He M. Anterior chamber angle assessment techniques // Survey of Ophthalmology. — 2008. — Vol. 53, no. 3. — P. 250–273.

[55] Radhakrishnan S., Goldsmith J., Huang D. et al. Comparison of optical coherence tomography and ultrasound biomicroscopy for detection of narrow anterior chamber angles // Archives of Ophthalmology. — 2005. — Vol. 123, no. 8. — P. 1053–1059.

[56] Konstantopoulos A., Hossain P., Anderson D. F. Recent advances in ophthalmic anterior segment imaging: a new era for ophthalmic diagnosis? // British Journal of Ophthalmology. — 2007. — Vol. 91, no. 4. — P. 551–557.

[57] Wojtkowski M., Kowalczyk A., Leitgeb R., Fercher A. Full range complex spectral optical coherence tomography technique in eye imaging // Optics Letters. — 2002. — Vol. 27, no. 16. — P. 1415–1417.

[58] Grulkowski I., Gora M., Szkulmowski M. et al. Anterior segment imaging with Spectral OCT system using a high-speed CMOS camera // Optics Express. — 2009. — Vol. 17, no. 6. — P. 4842–4858.

[59] Pérez-Merino P., Velasco-Ocana M., Martinez-Enriquez E., Marcos S. OCT-based crystalline lens topography in accommodating eyes // Biomed. Optics Express. — 2015. — Vol. 6, no. 12. — P. 5039–5054.

[60] Tian J., Marziliano P., Baskaran M. et al. Automatic anterior chamber angle assessment for HD-OCT images // IEEE Trans. on Biomed. Eng. — 2011. — Vol. 58, no. 11. — P. 3242–3249.

[61] Console J. W., Sakata L. M., Aung T. et al. Quantitative analysis of anterior segment optical coherence tomography images: the Zhongshan Angle Assessment Program // British Journal of Ophthalmology. — 2008. — Vol. 92, no. 12. — P. 1612–1616.

[62] Leung C. K.-s., Yung W.-h., Yiu C. K.-f. et al. Novel approach for anterior chamber angle analysis: anterior chamber angle detection with edge measurement and identification algorithm (ACADEMIA) // Archives of Ophthalmology. — 2006. — Vol. 124, no. 10. — P. 1395–1401.

[63] Sakata L. M., Lavanya R., Friedman D. S. et al. Assessment of the scleral spur in anterior segment optical coherence tomography images // Archives of Ophthalmology. — 2008. — Vol. 126, no. 2. — P. 181–185.

[64] Wu W., Li Y., Huang D., Duan H. A compound segmentation algorithm for anterior chamber angle in OCT image // 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI) / IEEE. — Vol. 1. — 2011. — P. 12–15.

[65] Niwas S. I., Lin W., Kwoh C. K. et al. Cross-examination for angle-closure glaucoma feature detection // IEEE Journal of Biomed. and Health Informatics. — 2016. — Vol. 20, no. 1. — P. 343–354.

[66] Niwas S. I., Lin W., Bai X. et al. Automated anterior segment OCT image analysis for Angle Closure Glaucoma mechanisms classification // Computer Methods and Programs in Biomed. — 2016. — Vol. 130. — P. 65–75.

[67] Xu Y., Liu J., Tan N. M. et al. Anterior chamber angle classification using multiscale histograms of oriented gradients for glaucoma subtype identification // 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society / IEEE. — 2012. — P. 3167–3170.

[68] IMO. Instituto de microcirugía ocular. — 2018. — Access mode: https://www.imo.es/en (online; accessed: 2018-01-03).

[69] Cha S.-H. Comprehensive survey on distance/similarity measures between probability density functions // International Journal of Mathematical Models and Methods in Applied Sciences. — 2007. — Vol. 1, no. 4. — P. 300–307.

[70] Van Der Maaten L., Postma E., Van den Herik J. Dimensionality reduction: a comparative // J Mach Learn Res. — 2009. — Vol. 10, no. 66-71. — P. 13.

[71] Grubbs F. E. Procedures for detecting outlying observations in samples // Technometrics. — 1969. — Vol. 11, no. 1. — P. 1–21.

[72] Hodge V., Austin J. A survey of outlier detection methodologies // Artificial intelligence review. — 2004. — Vol. 22, no. 2. — P. 85–126.

[73] Onorato M., Residori S., Bortolozzo U. et al. Rogue waves and their generating mechanisms in different physical contexts // Physics Reports. — 2013. — Vol. 528, no. 2. — P. 47–89.

[74] Solli D., Ropers C., Koonath P., Jalali B. Optical rogue waves // Nature. — 2007. — Vol. 450, no. 7172. — P. 1054–1057.

[75] Zhen-Ya Y. Financial rogue waves // Communications in Theoretical Physics. — 2010. — Vol. 54, no. 5. — P. 947.

[76] Shats M., Punzmann H., Xia H. Capillary rogue waves // Physical review letters. — 2010. — Vol. 104, no. 10. — P. 104503.

[77] Katz R. W., Parlange M. B., Naveau P. Statistics of extremes in hydrology // Advances in water resources. — 2002. — Vol. 25, no. 8-12. — P. 1287–1304.

[78] Chabchoub A., Hoffmann N., Akhmediev N. Rogue wave observation in a water wave tank // Physical Review Letters. — 2011. — Vol. 106, no. 20. — P. 204502.

[79] Akhmediev N., Kibler B., Baronio F. et al. Roadmap on optical rogue waves and extreme events // Journal of Optics. — 2016. — Vol. 18, no. 6. — P. 063001.

[80] Liu H., Shah S., Jiang W. On-line outlier detection and data cleaning // Computers & chemical engineering. — 2004. — Vol. 28, no. 9. — P. 1635–1647.

[81] Brodley C. E., Friedl M. A. et al. Identifying and eliminating mislabeled training instances // Proceedings of the National Conference on Artificial Intelligence. — 1996. — P. 799–805.

[82] Brodley C. E., Friedl M. A. Identifying mislabeled training data // Journal of artificial intelligence research. — 1999. — Vol. 11. — P. 131–167.

[83] Aleskerov E., Freisleben B., Rao B. Cardwatch: A neural network based database mining system for credit card fraud detection // Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr) / IEEE. — 1997. — P. 220–226.

[84] Cheng Q., Varshney P. K., Michels J. H., Belcastro C. M. Fault detection in dynamic systems via decision fusion // IEEE Transactions on aerospace and electronic systems. — 2008. — Vol. 44, no. 1. — P. 227–242.

[85] Pimentel M. A., Clifton D. A., Clifton L., Tarassenko L. A review of novelty detection // Signal Processing. — 2014. — Vol. 99. — P. 215–249.

[86] Agrawal S., Agrawal J. Survey on anomaly detection using data mining techniques // Procedia Computer Science. — 2015. — Vol. 60. — P. 708–713.

[87] Kou Y., Lu C.-T., Chen D. Spatial weighted outlier detection // Proceedings of the 2006 SIAM international conference on data mining / SIAM. — 2006. — P. 614–618.

[88] Lu C.-T., Chen D., Kou Y. Detecting spatial outliers with multiple attributes // Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence / IEEE. — 2003. — P. 122–128.

[89] Sun P., Chawla S. On local spatial outliers // Fourth IEEE International Conference on Data Mining (ICDM'04) / IEEE. — 2004. — P. 209–216.

[90] Spence C., Parra L., Sajda P. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model // Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001) / IEEE. — 2001. — P. 3–10.

[91] Taoum A., Mourad-Chehade F., Amoud H. Early-warning of ARDS using novelty detection and data fusion // Computers in biology and medicine. — 2018. — Vol. 102. — P. 191–199.

[92] Schlegl T., Seeböck P., Waldstein S. M. et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks // Medical image analysis. — 2019. — Vol. 54. — P. 30–44.

[93] Chandola V., Banerjee A., Kumar V. Anomaly detection for discrete sequences: A survey // IEEE Transactions on Knowledge and Data Engineering. — 2012. — Vol. 24, no. 5. — P. 823–839.

[94] Hawkins S., He H., Williams G., Baxter R. Outlier detection using replicator neural networks // International Conference on Data Warehousing and Knowledge Discovery / Springer. — 2002. — P. 170–180.

[95] Chen J., Sathe S., Aggarwal C., Turaga D. Outlier detection with autoencoder ensembles // Proceedings of the 2017 SIAM International Conference on Data Mining / SIAM. — 2017. — P. 90–98.

[96] Sabokrou M., Fayyaz M., Fathy M. et al. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes // Computer Vision and Image Understanding. — 2018. — Vol. 172. — P. 88–97.

[97] Zimek A., Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2018. — Vol. 8, no. 6. — P. e1280.

[98] Knox E. M., Ng R. T. Algorithms for mining distance-based outliers in large datasets // Proceedings of the international conference on very large data bases / Citeseer. — 1998. — P. 392–403.

[99] Ramaswamy S., Rastogi R., Shim K. Efficient algorithms for mining outliers from large data sets // ACM Sigmod Record / ACM. — Vol. 29. — 2000. — P. 427–438.

[100] Angiulli F., Pizzuti C. Outlier mining in large high-dimensional data sets // IEEE transactions on Knowledge and Data engineering. — 2005. — Vol. 17, no. 2. — P. 203–215.

[101] Angiulli F., Fassetti F. DOLPHIN: An efficient algorithm for mining distance-based outliers in very large datasets // ACM Transactions on Knowledge Discovery from Data (TKDD). — 2009. — Vol. 3, no. 1. — P. 4.

[102] Yang Z., Wu C., Chen T. et al. Detecting outlier measurements based on graph rigidity for wireless sensor network localization // IEEE Transactions on Vehicular Technology. — 2012. — Vol. 62, no. 1. — P. 374–383.

[103] Abukhalaf H., Wang J., Zhang S. Mobile-assisted anchor outlier detection for localization in wireless sensor networks // International Journal of Future Generation Communication and Networking. — 2016. — Vol. 9, no. 7. — P. 63–76.

[104] Abukhalaf H., Wang J., Zhang S. Outlier detection techniques for localization in wireless sensor networks: A survey // International Journal of Future Generation Communication and Networking. — 2015. — Vol. 8, no. 6. — P. 99–114.

[105] Pang Y., Yuan Y. Outlier-resisting graph embedding // Neurocomputing. — 2010. — Vol. 73, no. 4-6. — P. 968–974.

[106] Schubert E., Gertz M. Intrinsic t-stochastic neighbor embedding for visualization and outlier detection // International Conference on Similarity Search and Applications / Springer. — 2017. — P. 188–203.

[107] Madabhushi A., Shi J., Rosen M. et al. Graph embedding to improve supervised classification and novel class detection: application to prostate cancer // International Conference on Medical Image Computing and Computer-Assisted Intervention / Springer. — 2005. — P. 729–737.

[108] Cook D. J., Holder L. B. Graph-based data mining // IEEE Intelligent Systems and Their Applications. — 2000. — Vol. 15, no. 2. — P. 32–41.

[109] Eberle W., Holder L. Anomaly detection in data represented as graphs // Intelligent Data Analysis. — 2007. — Vol. 11, no. 6. — P. 663–689.

[110] Rahmani A., Afra S., Zarour O. et al. Graph-based approach for outlier detection in sequential data and its application on stock market and weather data // Knowledge-Based Systems. — 2014. — Vol. 61. — P. 89–97.

[111] Agovic A., Banerjee A., Ganguly A., Protopopescu V. Anomaly detection using manifold embedding and its applications in transportation corridors // Intelligent Data Analysis. — 2009. — Vol. 13, no. 3. — P. 435–455.

[112] Wang L., Li Z., Sun J. Improved IsoMap algorithm for anomaly detection in hyperspectral images // Fourth International Conference on Machine Vision (ICMV 2011): Machine Vision, Image Processing, and Pattern Analysis / International Society for Optics and Photonics. — Vol. 8349. — 2012. — P. 834902.

[113] Brito M., Chavez E., Quiroz A., Yukich J. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection // Statistics & Probability Letters. — 1997. — Vol. 35, no. 1. — P. 33–42.

[114] Barrat A., Barthélemy M., Vespignani A. Dynamical Processes on Complex Networks. — Cambridge University Press, 2008. — ISBN: 0521879507.

[115] Cohen R., Havlin S. Complex Networks: Structure, Robustness and Function. — Cambridge University Press, 2010. — ISBN: 0521841569.

[116] Stauffer D. Introduction To Percolation Theory: Revised Second Edition. — Taylor & Francis, 1994. — ISBN: 0748402535.

[117] Callaway D. S., Newman M. E. J., Strogatz S. H., Watts D. J. Network robustness and fragility: Percolation on random graphs // Physical Review Letters. — 2000. — Vol. 85, no. 25. — P. 5468–5471.

[118] Newman M. E. J., Ziff R. M. Fast monte carlo algorithm for site or bond percolation // Physical Review E. — 2001. — Vol. 64, no. 1. — P. 016706.

[119] Schölkopf B., Platt J. C., Shawe-Taylor J. et al. Estimating the support of a high-dimensional distribution // Neural computation. — 2001. — Vol. 13, no. 7. — P. 1443–1471.

[120] Nayak J., Acharya R., Bhat P. S. et al. Automated diagnosis of glaucoma using digital fundus images // Journal of Medical Systems. — 2009. — Vol. 33, no. 5. — P. 337.

[121] Mookiah M. R. K., Acharya U. R., Lim C. M. et al. Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features // Knowledge-Based Systems. — 2012. — Vol. 33. — P. 73–82.

[122] Acharya U. R., Bhat S., Koh J. E. et al. A novel algorithm to detect glaucoma risk using texton and local configuration pattern features extracted from fundus images // Computers in Biology and Medicine. — 2017. — Vol. 88. — P. 72–83.

[123] Maheshwari S., Pachori R. B., Acharya U. R. Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images // IEEE Journal of Biomedical and Health Informatics. — 2017. — Vol. 21, no. 3. — P. 803–813.

[124] Raghavendra U., Fujita H., Bhandary S. V. et al. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images // Information Sciences. — 2018. — Vol. 441. — P. 41–49.

[125] Gardner G., Keating D., Williamson T. H., Elliott A. T. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool // British Journal of Ophthalmology. — 1996. — Vol. 80, no. 11. — P. 940–944.

[126] Walter T., Klein J.-C., Massin P., Erginay A. A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina // IEEE Transactions on Medical Imaging. — 2002. — Vol. 21, no. 10. — P. 1236–1243.

[127] Sinthanayothin C., Boyce J. F., Williamson T. H. et al. Automated detection of diabetic retinopathy on digital fundus images // Diabetic Medicine. — 2002. — Vol. 19, no. 2. — P. 105–112.

[128] Sopharak A., Uyyanonvara B., Barman S., Williamson T. H. Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods // Computerized Medical Imaging and Graphics. — 2008. — Vol. 32, no. 8. — P. 720–727.

[129] Abràmoff M. D., Reinhardt J. M., Russell S. R. et al. Automated early detection of diabetic retinopathy // Ophthalmology. — 2010. — Vol. 117, no. 6. — P. 1147–1154.

[130] Gulshan V., Peng L., Coram M. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs // Jama. — 2016. — Vol. 316, no. 22. — P. 2402–2410.

[131] Mane V. M., Jadhav D. Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images // Biomedical Engineering/Biomedizinische Technik. — 2016. — Vol. 62, no. 3. — P. 321–332.

[132] Soomro T. A., Gao J., Khan T. et al. Computerised approaches for the detection of diabetic retinopathy using retinal fundus images: a survey // Pattern Analysis and Applications. — 2017. — Vol. 20, no. 4. — P. 927–961.

[133] Budak U., Şengür A., Guo Y., Akbulut Y. A novel microaneurysms detection approach based on convolutional neural networks with reinforcement sample learning algorithm // Health information science and systems. — 2017. — Vol. 5, no. 1. — P. 14.

[134] Hussain M. A., Bhuiyan A., Luu C. D. et al. Classification of healthy and diseased retina using SD-OCT imaging and random forest algorithm // PloS one. — 2018. — Vol. 13, no. 6. — P. e0198281.

[135] Koh J. E., Ng E. Y., Bhandary S. V. et al. Automated detection of retinal health using PHOG and SURF features extracted from fundus images // Applied Intelligence. — 2017. — Vol. 48, no. 5. — P. 1379–1393.

[136] Abràmoff M. D., Garvin M. K., Sonka M. Retinal imaging and image analysis // IEEE reviews in biomedical engineering. — 2010. — Vol. 3. — P. 169–208.

[137] Sun C., Wang J. J., Mackey D. A., Wong T. Y. Retinal vascular caliber: systemic, environmental, and genetic associations // Survey of ophthalmology. — 2009. — Vol. 54, no. 1. — P. 74–95.

[138] Cheung C. Y.-l., Ong Y. T., Ikram M. K. et al. Microvascular network alterations in the retina of patients with alzheimer's disease // Alzheimer's & Dementia: the Journal of the Alzheimer's Association. — 2014. — Vol. 10, no. 2. — P. 135–142.

[139] Chan V., Tso T., Tang F. et al. Using retinal imaging to study dementia // Journal of Visualized Experiments: JoVE. — 2017. — no. 129.

[140] Yin X., Ng B. W., He J. et al. Accurate image analysis of the retina using hessian matrix and binarisation of thresholded entropy with application of texture mapping // PLoS One. — 2014. — Vol. 9, no. 4. — P. e95943.

[141] Memari N., Ramli A. R., Saripan M. I. B. et al. Supervised retinal vessel segmentation from color fundus images based on matched filtering and AdaBoost classifier // PloS one. — 2017. — Vol. 12, no. 12. — P. e0188939.

[142] Hoover A., Kouznetsova V., Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response // IEEE Transactions on Medical Imaging. — 2000. — Vol. 19, no. 3. — P. 203–210.

[143] Fraz M. M., Remagnino P., Hoppe A. et al. Blood vessel segmentation methodologies in retinal images–a survey // Computer Methods and Programs in Biomedicine. — 2012. — Vol. 108, no. 1. — P. 407–433.

[144] Almotiri J., Elleithy K., Elleithy A. Retinal vessels segmentation techniques and algorithms: A survey // Applied Sciences. — 2018. — Vol. 8, no. 2. — P. 155.

[145] Moccia S., De Momi E., El Hadji S., Mattos L. S. Blood vessel segmentation algorithms-review of methods, datasets and evaluation metrics // Computer Methods and Programs in Biomedicine. — 2018. — Vol. 158. — P. 71–91.

[146] Guo Y., Budak Ü., Şengür A. A novel retinal vessel detection approach based on multiple deep convolution neural networks // Computer methods and programs in biomedicine. — 2018. — Vol. 167. — P. 43–48.

[147] Guo Y., Budak Ü., Vespa L. J. et al. A retinal vessel detection approach using convolution neural network with reinforcement sample learning strategy // Measurement. — 2018. — Vol. 125. — P. 586–591.

[148] Guo Y., Budak Ü., Şengür A., Smarandache F. A retinal vessel detection approach based on shearlet transform and indeterminacy filtering on fundus images // Symmetry. — 2017. — Vol. 9, no. 10. — P. 235.

[149] Mandelbrot B. B. The fractal geometry of nature. — WH freeman New York, 1983. — ISBN: 0910321647.

[150] Family F., Masters B. R., Platt D. E. Fractal pattern formation in human retinal vessels // Physica D: Nonlinear Phenomena. — 1989. — Vol. 38, no. 1-3. — P. 98–103.

[151] Masters B. R. Fractal analysis of the vascular tree in the human retina // Annu. Rev. Biomed. Eng. — 2004. — Vol. 6. — P. 427–452.

[152] Orlando J. I., Van Keer K., Barbosa Breda J. et al. Proliferative diabetic retinopathy characterization based on fractal features: Evaluation on a publicly available dataset // Medical Physics. — 2017. — Vol. 44, no. 12. — P. 6425–6434.

[153] Huang F., Dashtbozorg B., Zhang J. et al. Reliability of using retinal vascular fractal dimension as a biomarker in the diabetic retinopathy detection // Journal of Ophthalmology. — 2016. — Vol. 2016.

[154] Stosic T., Stosic B. D. Multifractal analysis of human retinal vessels // IEEE Transactions on Medical Imaging. — 2006. — Vol. 25, no. 8. — P. 1101–1107.

[155] Ţălu Ş., Stach S., Călugăru D. M. et al. Analysis of normal human retinal vascular network architecture using multifractal geometry // International Journal of Ophthalmology. — 2017. — Vol. 10, no. 3. — P. 434.

[156] Azemin M. C., Kumar D. K., Wong T. Y. et al. Robust methodology for fractal analysis of the retinal vasculature // IEEE Transactions on Medical Imaging. — 2011. — Vol. 30, no. 2. — P. 243–250.

[157] Colomer A., Naranjo V., Janvier T., Mossi J. M. Evaluation of fractal dimension effectiveness for damage detection in retinal background // Journal of Computational and Applied Mathematics. — 2018. — Vol. 337. — P. 341–353.

[158] Albert R., Barabási A.-L. Statistical mechanics of complex networks // Reviews of modern physics. — 2002. — Vol. 74, no. 1. — P. 47.

[159] Boccaletti S., Latora V., Moreno Y. et al. Complex networks: Structure and dynamics // Physics reports. — 2006. — Vol. 424, no. 4-5. — P. 175–308.

[160] Schieber T. A., Carpi L., Díaz-Guilera A. et al. Quantification of network structural dissimilarities // Nature Communications. — 2017. — Vol. 8. — P. 13928.

[161] Van Wijk B. C., Stam C. J., Daffertshofer A. Comparing brain networks of different size and connectivity density using graph theory // PloS One. — 2010. — Vol. 5, no. 10. — P. e13701.

[162] Martínez J. H., Chavez M. Comparing complex networks: in defence of the simple // New Journal of Physics. — 2019. — Vol. 21, no. 1. — P. 013033.

[163] Liebovitch L. S., Toth T. A fast algorithm to determine fractal dimensions by box counting // Physics Letters A. — 1989. — Vol. 141, no. 8-9. — P. 386–390.

[164] Mudigonda S., Oloumi F., Katta K. M., Rangayyan R. M. Fractal analysis of neovascularization due to diabetic retinopathy in retinal fundus images // E-Health and Bioengineering Conference (EHB), 2015 / IEEE. — 2015. — P. 1–4.

[165] Akram M. U., Tariq A., Khalid S. et al. Glaucoma detection using novel optic disc localization, hybrid feature set and classification techniques // Australasian Physical & Engineering Sciences in Medicine. — 2015. — Vol. 38, no. 4. — P. 643–655.

[166] Estrada R., Allingham M. J., Mettu P. S. et al. Retinal artery-vein classification via topology estimation // IEEE transactions on medical imaging. — 2015. — Vol. 34, no. 12. — P. 2518–2534.

[167] Girard F., Kavalec C., Cheriet F. Joint segmentation and classification of retinal arteries/veins from fundus images // Artificial intelligence in medicine. — 2019. — Vol. 94. — P. 96–109.

[168] Aibinu A. M., Iqbal M. I., Shafie A. A. et al. Vascular intersection detection in retina fundus images using a new hybrid approach // Computers in Biology and Medicine. — 2010. — Vol. 40, no. 1. — P. 81–89.

[169] Calvo D., Ortega M., Penedo M. G., Rouco J. Automatic detection and characterisation of retinal vessel tree bifurcations and crossovers in eye fundus images // Computer methods and programs in biomedicine. — 2011. — Vol. 103, no. 1. — P. 28–38.

[170] Bhuiyan A., Nath B., Chua J., Ramamohanarao K. Automatic detection of vascular bifurcations and crossovers from color retinal fundus images // 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System / IEEE. — 2007. — P. 711–718.

[171] Garside K., Henderson R., Makarenko I., Masoller C. Topological data analysis of high resolution diabetic retinopathy images // PloS one. — 2019. — Vol. 14, no. 5. — P. e0217413.

[172] Wieczorek S., Krauskopf B., Simpson T. B., Lenstra D. The dynamical complexity of optically injected semiconductor lasers // Physics Reports. — 2005. — Vol. 416, no. 1-2. — P. 1–128.

[173] Lau E. K., Zhao X., Sung H.-K. et al. Strong optical injection-locked semiconductor lasers demonstrating >100-GHz resonance frequencies and 80-GHz intrinsic bandwidths // Optics Express. — 2008. — Vol. 16, no. 9. — P. 6609–6618.

[174] Lo K.-H., Hwang S.-K., Donati S. Numerical study of ultrashort-optical-feedback-enhanced photonic microwave generation using optically injected semiconductor lasers at period-one nonlinear dynamics // Optics express. — 2017. — Vol. 25, no. 25. — P. 31595–31611.

[175] Xue C., Ji S., Wang A. et al. Narrow-linewidth single-frequency photonic microwave generation in optically injected semiconductor lasers with filtered optical feedback // Optics letters. — 2018. — Vol. 43, no. 17. — P. 4184–4187.

[176] Li X.-Z., Chan S.-C. Heterodyne random bit generation using an optically injected semiconductor laser in chaos // IEEE Journal of Quantum Electronics. — 2013. — Vol. 49, no. 10. — P. 829–838.

[177] Köllisch N., Behrendt J., Klein M., Hoffmann N. Nonlinear real time prediction of ocean surface waves // Ocean Engineering. — 2018. — Vol. 157. — P. 387–400.

[178] Franzke C. Predictability of extreme events in a nonlinear stochastic-dynamical model // Physical Review E. — 2012. — Vol. 85, no. 3. — P. 031134.

[179] Birkholz S., Brée C., Demircan A., Steinmeyer G. Predictability of rogue events // Physical review letters. — 2015. — Vol. 114, no. 21. — P. 213901.

[180] Pathak J., Hunt B., Girvan M. et al. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach // Physical review letters. — 2018. — Vol. 120, no. 2. — P. 024102.

[181] Isensee J., Datseris G., Parlitz U. Predicting spatio-temporal time series using dimension reduced local states // Journal of Nonlinear Science. — 2019. — Oct. — P. 1–23.

[182] Bialonski S., Ansmann G., Kantz H. Data-driven prediction and prevention of extreme events in a spatially extended excitable system // Physical Review E. — 2015. — Vol. 92, no. 4. — P. 042910.

[183] Kuremoto T., Kimura S., Kobayashi K., Obayashi M. Time series forecasting using a deep belief network with restricted boltzmann machines // Neurocomputing. — 2014. — Vol. 137. — P. 47–56.

[184] Wang J., Chi D., Wu J., Lu H.-y. Chaotic time series method combined with particle swarm optimization and trend adjustment for electricity demand forecasting // Expert Systems with Applications. — 2011. — Vol. 38, no. 7. — P. 8419–8429.

[185] Ardalani-Farsa M., Zolfaghari S. Chaotic time series prediction with residual analysis method using hybrid Elman–NARX neural networks // Neurocomputing. — 2010. — Vol. 73, no. 13-15. — P. 2540–2553.

[186] Gholipour A., Araabi B. N., Lucas C. Predicting chaotic time series using neural and neurofuzzy models: a comparative study // neural processing letters. — 2006. — Vol. 24, no. 3. — P. 217–239.

[187] Lau K., Wu Q. Local prediction of non-linear time series using support vector regression // Pattern Recognition. — 2008. — Vol. 41, no. 5. — P. 1539–1547.

[188] Bonatto C., Feyereisen M., Barland S. et al. Deterministic optical rogue waves // Physical review letters. — 2011. — Vol. 107, no. 5. — P. 053901.

[189] Cavalcante H. L. d. S., Oriá M., Sornette D. et al. Predictability and suppression of extreme events in a chaotic system // Physical review letters. — 2013. — Vol. 111, no. 19. — P. 198701.

[190] Zamora-Munt J., Garbin B., Barland S. et al. Rogue waves in optically injected lasers: Origin, predictability, and suppression // Physical Review A. — 2013. — Vol. 87, no. 3. — P. 035802.

[191] San Miguel M., Toral R. Stochastic effects in physical systems // Instabilities and nonequilibrium structures VI. — Springer, 2000. — P. 35–127.

[192] He Z., Wen X., Liu H., Du J. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region // Journal of Hydrology. — 2014. — Vol. 509. — P. 379–386.

[193] Altman N. S. An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician. — 1992. — Vol. 46, no. 3. — P. 175–185.

[194] Boser B. E., Guyon I. M., Vapnik V. N. A training algorithm for optimal margin classifiers // Proceedings of the fifth annual workshop on Computational learning theory / ACM. — 1992. — P. 144–152.

[195] Vapnik V. The nature of statistical learning theory. — Springer, New York, NY, 2010. — ISBN: 9781441931603.

[196] Huang T.-M., Kecman V., Kopriva I. Kernel based algorithms for mining huge data sets. — Springer, 2006. — Vol. 1. — ISBN: 3540316817.

[197] Drucker H., Burges C. J., Kaufman L. et al. Support vector regression machines // Advances in neural information processing systems. — 1997. — P. 155–161.

[198] LeCun Y., Boser B. E., Denker J. S. et al. Handwritten digit recognition with a back-propagation network // Advances in neural information processing systems. — 1990. — P. 396–404.

[199] Verstraeten D., Schrauwen B., D'Haene M., Stroobandt D. An experimental unification of reservoir computing methods // Neural Networks. — 2007. — Vol. 20, no. 3. — P. 391 – 403.

[200] Rodan A., Tino P. Minimum complexity echo state network // IEEE transactions on neural networks. — 2011. — Vol. 22, no. 1. — P. 131–144.

[201] Lukoševičius M., Jaeger H. Reservoir computing approaches to recurrent neural network training // Computer Science Review. — 2009. — Vol. 3, no. 3. — P. 127–149.

[202] Larger L., Soriano M. C., Brunner D. et al. Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing // Optics express. — 2012. — Vol. 20, no. 3. — P. 3241–3249.

[203] Paquot Y., Duport F., Smerieri A. et al. Optoelectronic reservoir computing // Scientific reports. — 2012. — Vol. 2. — P. 287.

[204] Kawai Y., Park J., Asada M. A small-world topology enhances the echo state property and signal propagation in reservoir computing // Neural Networks. — 2019. — Vol. 112. — P. 15–23.

[205] Griffith A., Pomerance A., Gauthier D. J. Forecasting chaotic systems with very low connectivity reservoir computers // arXiv:1910.00659. — 2019.

[206] Yang Z. O., Wang Y., Li D., Wang C. Predict the time series of the parameter-varying chaotic system based on reduced recursive lease square support vector machine // 2009 International Conference on Artificial Intelligence and Computational Intelligence / IEEE. — Vol. 1. — 2009. — P. 29–34.

[207] Alfaras M., Soriano M. C., Ortín S. A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection // Frontiers in Physics. — 2019. — Vol. 7. — P. 103.

[208] McLaren K. XIII-the development of the CIE 1976 (L* a* b*) uniform colour space and colour-difference formula // Journal of the Society of Dyers and Colourists. — 1976. — Vol. 92, no. 9. — P. 338–341.

[209] Lowell J., Hunter A., Steel D. et al. Optic nerve head segmentation // IEEE Transactions on Medical Imaging. — 2004. — Vol. 23, no. 2. — P. 256–264.

Appendix A.

Outlier mining methods on synthetic databases

In this appendix we show more results of the methods proposed in chapter 3 on different synthetic databases.

A.1. Points in a high dimensional Euclidean space

A.1.1. Ramaswamy-like (HS01)

We generated two sets in a similar manner as those generated in Ref. [99], and used the first one to train the parameters of the methods that needed training, and the second one to evaluate the results. We generated points in a 15-dimensional space, normal points were distributed in 25 hyper-spherical clusters whose centers were arranged in a regular grid (of 5 by 5) in the 2-dimensional plane varying the first 2 dimensions (while kipping the other 13 at zero). While the outliers were randomly sampled in a hypercube. We used 100 points per cluster (2500 normal points) and 50 outliers. We show the results for this database in Fig. A.1.1.

A.1.2. N-D Spherical cap (SC05)

We generated two sets of points in a 5-dimensional space, and used the first one to train the parameters of the methods that needed training, and the second one to evaluate the results. Normal points were randomly generated lying in 5-dimensional hyper-spherical cap, while outliers were

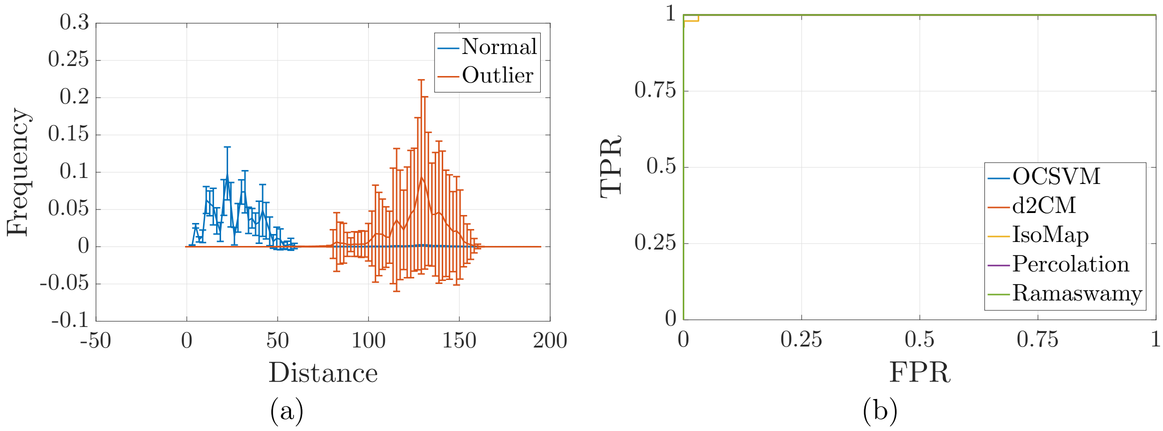


Figure A.1.1.: Results for the the HS01 database. (a) Distance histogram for normal and outlying points. (b) ROC curves for the 5 studied methods.

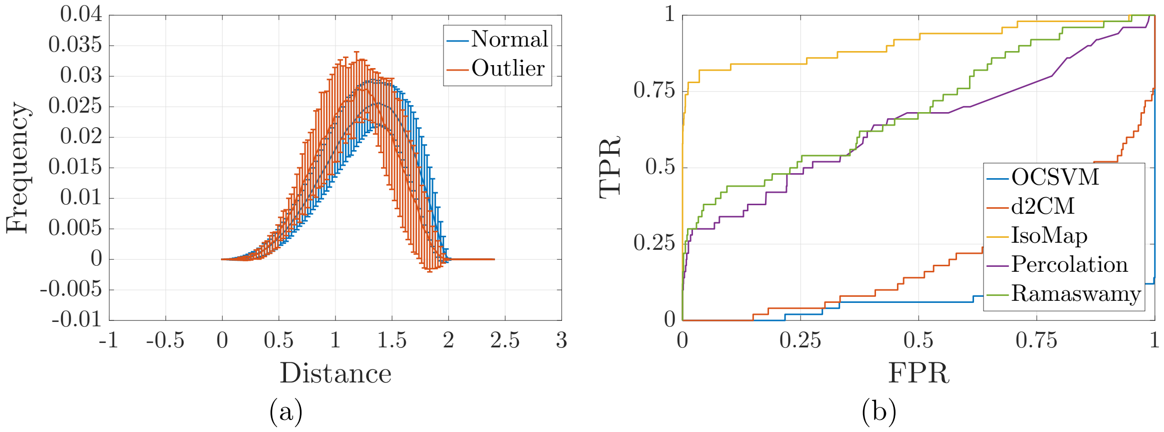


Figure A.1.2.: Results for the the SC05 database. (a) Distance histogram for normal and outlying points. (b) ROC curves for the 5 studied methods.

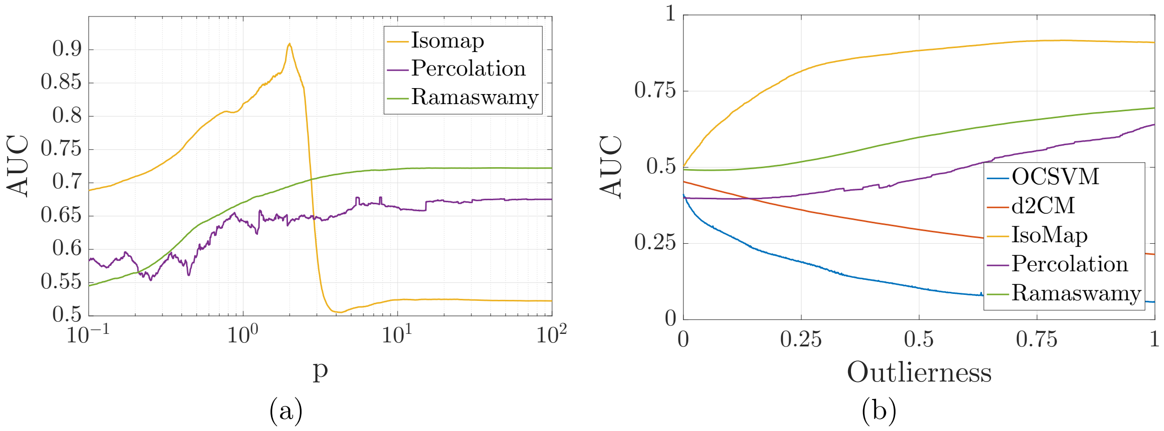


Figure A.1.3.: Robustness study for SC05 database. (a) Area under ROC curve as a function of p. (b) Area under ROC curve as a function of the outlierness.

generated randomly inside the hyper-sphere. We used 2000 normal points and 50 outliers. We show the results for this database in Fig. A.1.2.

With this database we also studied how robust the methods are when changing the distance measure [Fig. A.1.3.(a)] and the degree of "outlierness" [Fig. A.1.3.(b)]. To see how sensitive are the methods to changes on the distance measure, we tested the methods that use a distance in the second set varying p in the Minkowski distance family with the parameters that were set with $p = 2$ (Euclidean) using the first set. To see how sensitive the methods are to changes in the "outlierness" we modified the second set according to an outlierness level, when this outlierness is 1, no change was done on the set, when it decreases, the outliers were linearly moved to the spherical cap, and when the level equals 0, all outliers are perfectly projected on the spherical cap (so they are indistinguishable from the normal points).

A.2. Binary triangles

These databases were both generated in a similar manner. Binary images with resolution of 768 by 1024 pixels depicting equilateral triangles with added random noise (see example in Fig. A.2.1) were used as items in the database. The position, orientation, and size of the triangles was set

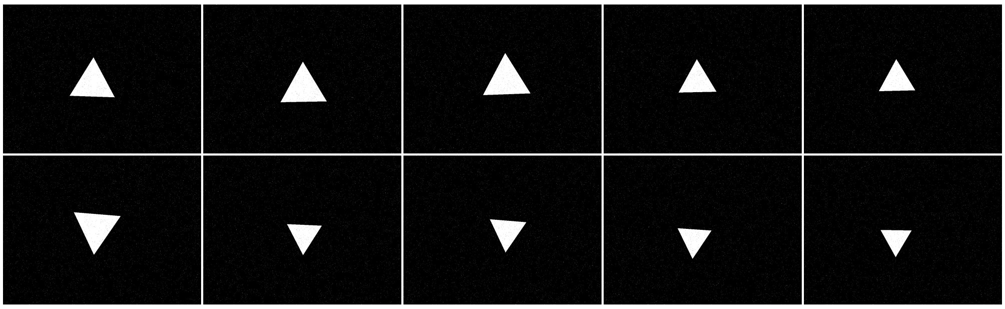


Figure A.2.1.: Example images from the database TR01. 5 normal examples (top) and 5 outlier examples (bottom).

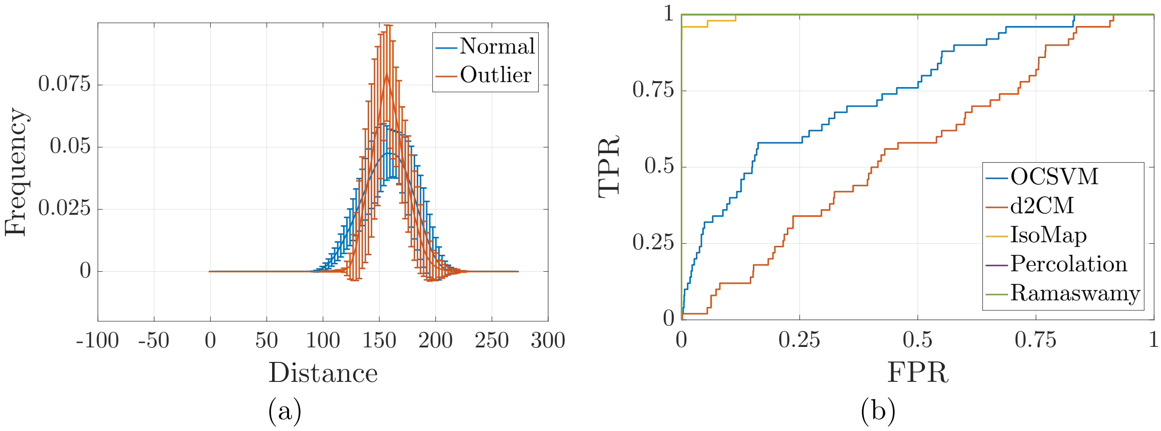


Figure A.2.2.: Results for the the TR01 database. (a) Distance histogram for normal and outlying points. (b) ROC curves for the 5 studied methods.

by random variables whose distribution was varied from database to database. In both cases we generated two sets with the same parameters and used the first one to train the parameters of the methods that needed training, and the second one to evaluate the results.

A.2.1. TR01

In this database we set the position of all triangles at the center (with some random variability) and all with the same size (also with some random variability), the only difference between normal and outlier triangles was the orientation (see Table A.1).

	Normal	Outliers
X position	512±50	512±50
Y position	384±50	384±50
Size	200±50	200±50
Orientation	0±5	60±5
Noise probability	0.5%	0.5%

Table A.1.: Parameters for generating the TR01 database. All the variables were sampled from a uniform distribution.

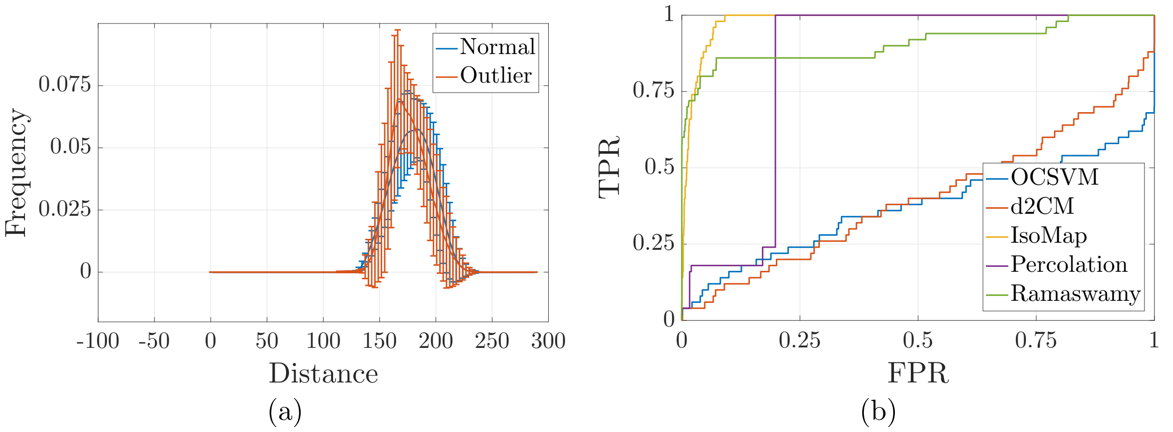


Figure A.2.3.: Results for the the TR02 database. (a) Distance histogram for normal and outlying points. (b) ROC curves for the 5 studied methods.

	Normal	Outliers
X position	512±50	512±70
Y position	384±50	384±70
Size	200±50	200±50
Orientation	0±5	60±20
Noise probability	1%	0.5%

Table A.2.: Parameters for generating the TR02 database. All the variables were sampled from a uniform distribution.

A.2.2. TR02

In this database we repeated the same study but increased the noise only in the normal images (see Table A.2). This was done because in the previous database, the distance distribution of normal and outlying points have regions in which they do not overlap (see Fig. A.2.2.a distances around 110), which was probably the reason of such a good performance for the distance-based methods. We set the noise probability in this database such that there exists overlap in all the bins of the distance histogram (see Fig. A.2.3.a)

For this database we also tested the robustness of the methods when changing the "outlierness" level. For that, we varied the orientation (tilt) of the outlying triangles from 0º (same as the normal ones) to 60º (opposite to the normal ones), see the results in Fig. A.2.4.

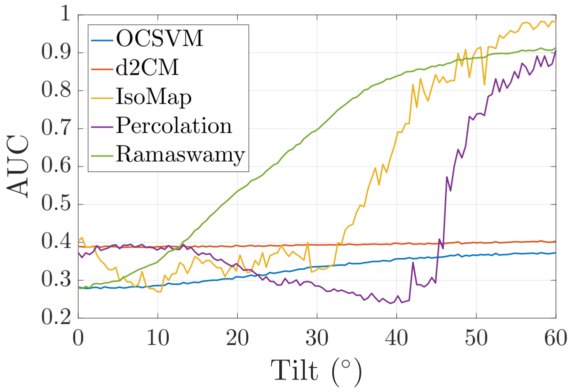


Figure A.2.4.: Robustness study for TR02 database. Area under ROC curve as a function tilt.

DataBase	OCSVM	d2CM	IsoMap	Percolation	Ramaswamy
HS01	1	1	1	1	1
HS01*	1	1	0.999	1	1
SC05	0.0847	0.21	0.958	0.775	0.79
SC05*	0.0576	0.214	0.91	0.64	0.695
TR01	0.764	0.574	0.999	1	1
TR01*	0.75	0.558	0.997	1	1
TR02	0.303	0.341	0.998	0.822	0.864
TR02*	0.372	0.402	0.981	0.837	0.908

Table A.3.: Area under ROC curves resulted for applying the described methods to the databases. Databases with an asterisk (*) were only used for testing.

A.3. Results Summary

We summarized the results of this appendix in Table A.3. Each database is shown two times, the first is the one used for training while the second one (marked with an *) was only used for testing.

Appendix B.

Fundus image segmentation and network information retrieval

In this appendix we describe the four steps in the image processing workflow employed for retrieving the topological information of the vessel network in retinal fundus images: preprocessing; segmentation; postprocessing; and network identification.

B.1. Preprocessing

In order to enhance the contrast between the vessel network and the retina, we apply a cascade of filters.

Each raw image (see Fig. B.1.1.(a)) is first equalized using a Contrast-limited Adaptive Histogram Equalization (Using MatLab function *adapthisteq*) separately to each color layer with 16-by-16 tiles and 256 bins, see Fig. B.1.1.(b).

Then, we perform a Gaussian high-pass filtering and retain the negative part alone of the resulting image. This way only the sudden drops of intensity are considered, see Fig. B.1.1.(c).

Next we apply a bank of 16 directional high-pass filters isotropically and compute the standard deviation pixel-by-pixel and color-by-color. This operation keeps only linear (one-dimensional) drops in the intensity of the original image (as vessels are) and disregards small points or areas (two-dimensional objects) with intensity drops.

The result is then masked with a simple mask computed by thresholding a blurred gray version of the original image.

An example of the resulting image after this preprocessing is shown in Fig. B.1.1.(d).

B.2. Segmentation

For the segmentation we adapted the algorithm proposed in Ref. [41] to work with color images. We define a metric between colors in the CIE 1976 L*a*b* color space [208] that amplifies the differences in global intensities and attenuates the differences in the red channel intensities (as the differences between vessels and retina are mainly on the green channel and the red channel has more background variations unrelated to the vessel network). The outcome is depicted in Fig. B.1.1.(e) where the background appears in black and the foreground mask, containing the vessel network information, is shown in white. As a comparison, the corresponding manual segmentation performed by a human expert is shown in Fig. B.1.1.(f).

We use a simplified version of the algorithm proposed in Ref. [209] for the optic disk localization to find and to segment the optical nerve, as it is shown in blue in Fig. B.1.1.(g).

B.3. Postprocessing & network identification

After the image segmentation step a series of morphological operations is performed to further clean up the segmented image. This information is merged with the optical nerve as depicted in Fig. B.1.1.(g). Finally, a skeletonization is performed (using the MatLab implementation in the *bwmorph* function) and the result is analyzed to retrieve the network structural information. The skeletonized version is topologically identical to the raw segmented mask, but the width of all vessels is set to one pixel. Thus, the pixels that have one neighbor are endpoints (we identify them as nodes), the pixels that have two neighbors are part of a vessel (we set them as part of a link connecting two nodes), and the pixels with three or more neighbors are bifurcation points (we identify them as nodes).

Therefore, we end up with a list of nodes (bifurcation points and endpoints) with their locations [marked in magenta in Fig. B.1.1.(h)], and the path of all the vessels connecting two nodes (colored in green). We define the whole optical nerve as a single node, allowing us to denote a central node and also disregard all possible vessel segmentation errors that may occur within the optical nerve. From the information plotted in Fig. B.1.1.(h) we can also retrieve an adjacency matrix, $\mathbf{A} = (A_{i,j})$, by assigning each node a natural number (which are the indices of the matrix) and setting $A_{i,j} = 1$ if nodes i and j are connected with a link and $A_{i,j} = 0$ otherwise. The length, $L_{i,j}$, and the width, $W_{i,j}$, of each link can be computed using the information contained in the skeletonized and raw segmented masks. The length accounts for the number of pixels spanned by each link in the skeletonized version while the width can be estimated from the number of pixels ($N_{i,j}$) each link has in the raw segmented mask as $N_{i,j} = L_{i,j} \times W_{i,j}$.

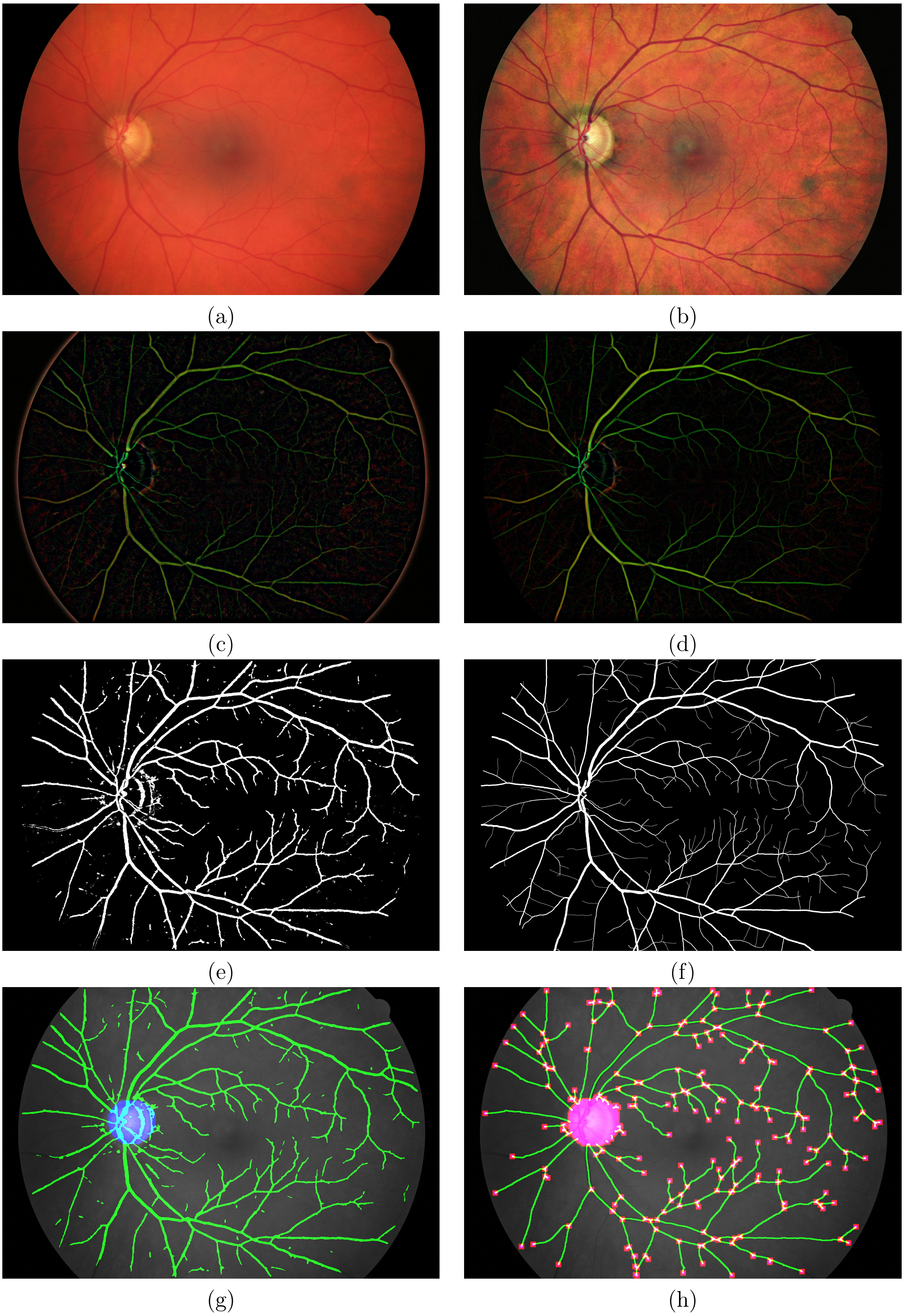
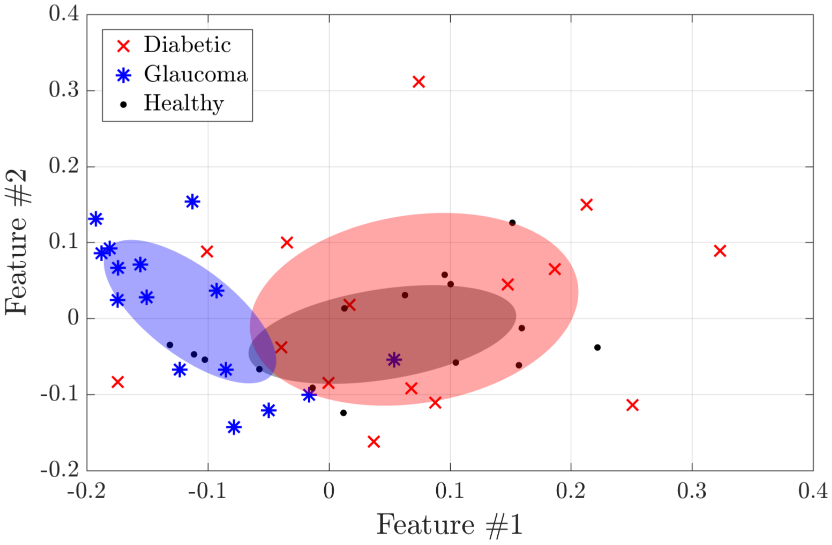


Figure B.1.1.: Image processing workflow. (a) Example image from the HRF database. (b) Image after the Contrast-limited Adaptive Histogram Equalization. (c) Gaussian high-pass filtering result. (d) Vessel enhancement result. Note that the arteries have a green color while veins have a slightly more yellowish color. (e) Raw segmentation result. (f) Manual segmentation performed by a human expert (for comparison). (g) Cleaned segmentation (in green), and optic nerve (in blue) superimposed to the gray-scale original image. (h) Network identification result, nodes in magenta and links in green superimposed to the gray-scale original image.

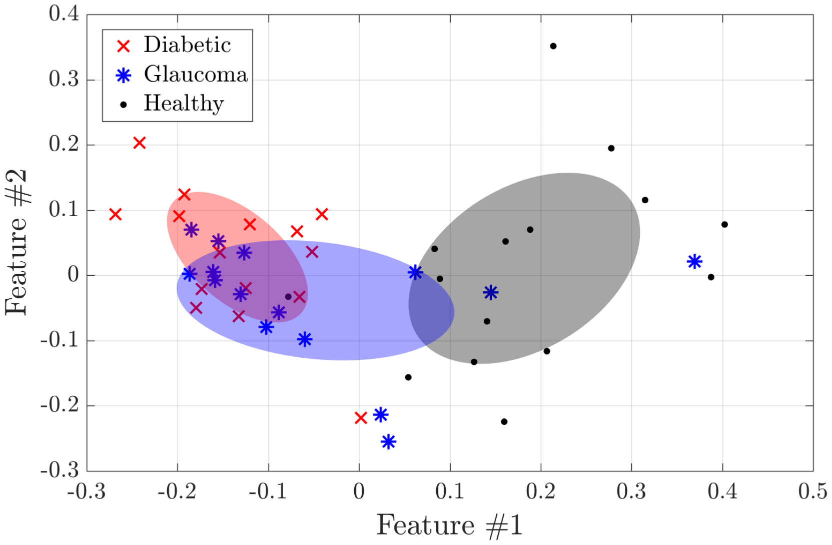
Appendix C.

Extra analyses of network-based features of the retina

In this appendix we show some extra figures with more examples of analyses made to the databases in chapter 4. In Fig. C.0.1 we show the results of analyzing the HRF database with the NND method using $l = 1$ and $a = 2$. In Fig. C.0.2 we show the results of analyzing the Messidor database with the WDD method using $l = 0$ and $a = 1$. Finally, in Figs. C.0.3 and C.0.4 we show, for all databases, the number of links *vs.* number of nodes.



(a)



(b)

Figure C.0.1.: IsoMap features for the Central NDD analysis of the HRF database with $l = 1$ and $a = 2$ for the (a) automated and (b) manual segmentations.

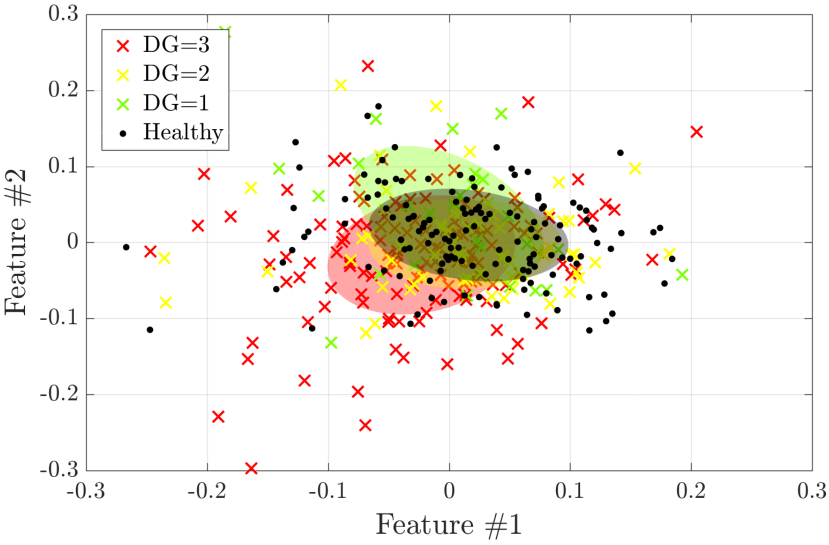
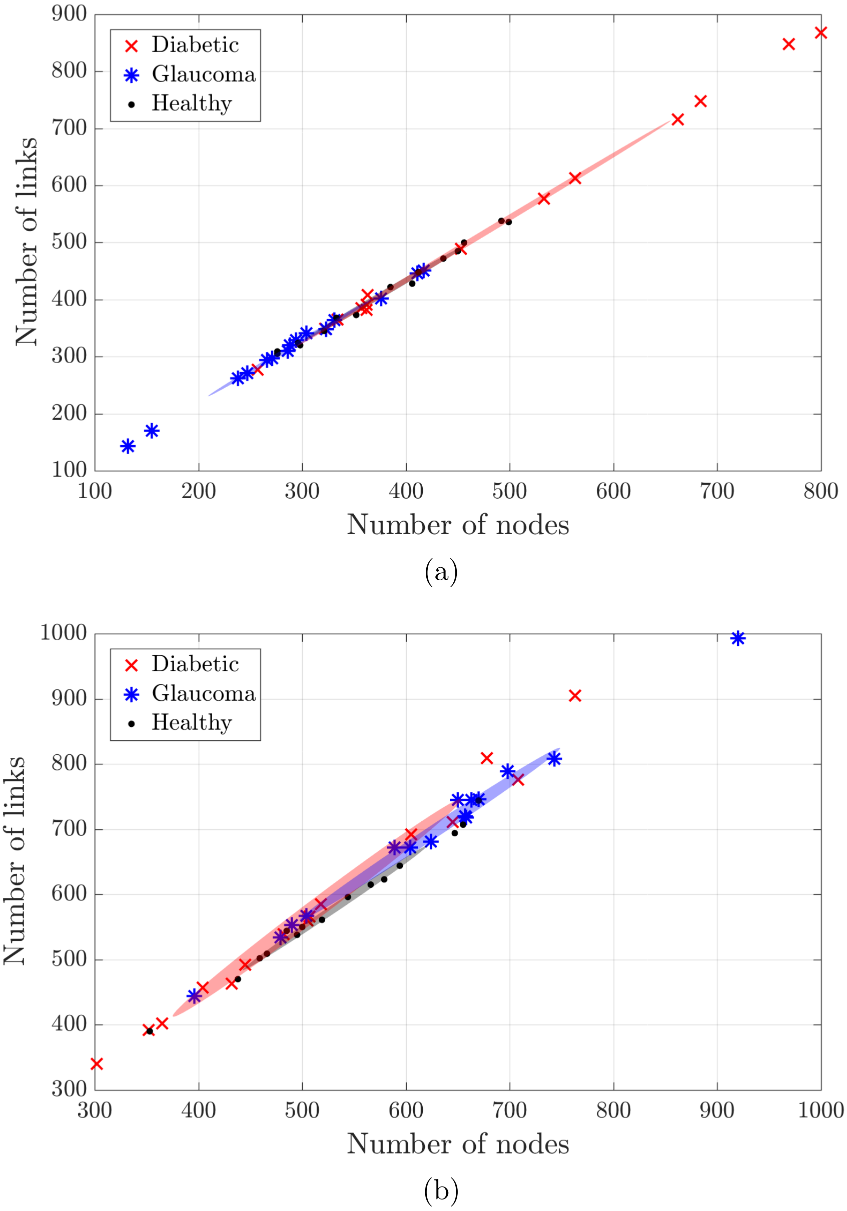


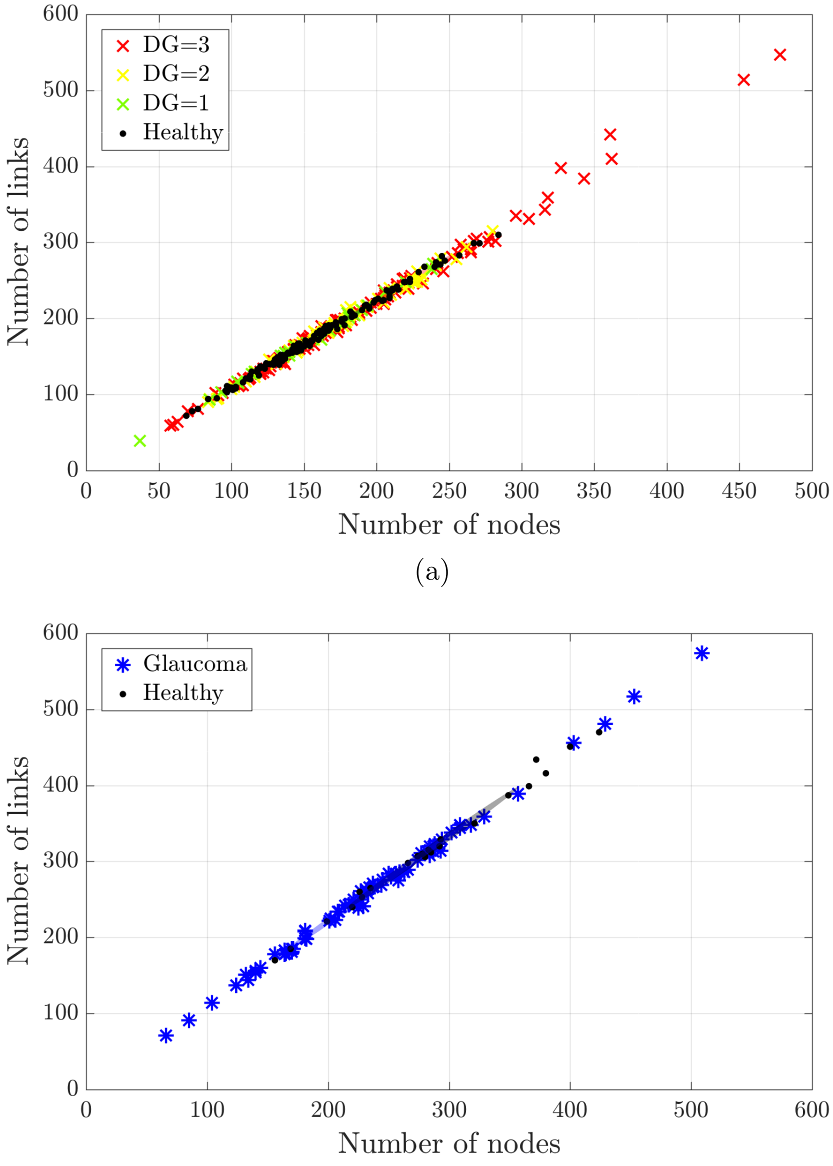
Figure C.0.2.: IsoMap features for the WDD analysis of the Messidor database with $l = 0$ and $a = 1$.



(a)

(b)

Figure C.0.3.: Number of links *vs.* number of nodes for the HRF database using the automated (a) and manual (b) segmentations.



(a)

(b)

Figure C.0.4.: Number of links *vs.* number of nodes for the automatic segmentations of Messidor (a) and IMO (b) databases.

Appendix D.

Publications and patents

D.1. Journal articles

D.1.1. Unsupervised feature extraction of anterior chamber OCT images for ordering and classification

Authors: Pablo Amil, Laura González, Elena Arrondo, Cecilia Salinas, J. L. Güell, Cristina Masoller, and Ulrich Parlitz

Journal: Scientific reports; year: 2019; volume: 9; page: 1157.

doi:10.1038/s41598-018-38136-8

Journal Quartile (2018): Q1; Journal Impact Factor (2018): 4.011

Citations to paper: 2

Reference: [7]

D.1.2. Outlier mining methods based on graph structure analysis

Authors: Pablo Amil, Nahuel Almeira, and Cristina Masoller

Journal: Frontiers in Physics; year: 2019; volume: 7; page: 194

doi:10.3389/fphy.2019.00194

Journal Quartile (2018): Q2; Journal Impact Factor (2018): 1.895

Citations to paper: 0

Reference: [8]

D.1.3. Network-based features for retinal fundus vessel structure analysis

Authors: Pablo Amil, Cesar F. Reyes-Manzano, Lev Guzmán-Vargas, Irene Sendiña-Nadal, and Cristina Masoller

Journal: PLOS ONE; year: 2019; volume: 14; page: e0220132

doi:10.1371/journal.pone.0220132

Journal Quartile (2018): Q2; Journal Impact Factor (2018): 2.776

Citations to paper: 0

Reference: [9]

D.1.4. Machine learning algorithms for predicting the amplitude of chaotic laser pulses

Authors: Pablo Amil, Miguel C. Soriano, and Cristina Masoller

Journal: Chaos; year: 2019; volume: 29; page: 113111

doi:10.1063/1.5120755

Journal Quartile (2018): Q1; Journal Impact Factor (2018): 2.643

Citations to paper: 0

Reference: [10]

D.2. Patents

D.2.1. Image processing method for glaucoma detection and computer program products thereof

Inventors: Amil Marletti, Pablo; Masoller Alonso, Cristina; Arrondo Murillo, Elena; Parlitz, Ulrich; Salinas Almela, Cecilia

Number: WO/2019/116074; Date: 2017-12-11; Country: Spain

D.2.2. A computer implemented method, a system and computer programs for anomaly detection using network analysis

Inventors: Amil Marletti, Pablo; Almeira, Nahuel; Masoller Alonso, Cristina

Number: EPC-19382388.7-1218; Date: 2019-05-17; Country: Spain

Appendix E.

Activities

E.1. Research stays

E.1.1. Center for Biomedical Technology

Date:12/12/2016

Responsible: Irene Sendiña

Place: Madrid, Spain

E.1.2. Max Planck Institute for Dynamics and Self-Organization

Dates:(06/03-01/05)/2017

Responsible: Ulrich Parlitz

Place: Göttingen, Germany

E.1.3. Max Planck Institute for Dynamics and Self-Organization

Dates:(30/07-16/08)/2017

Responsible: Ulrich Parlitz

Place: Göttingen, Germany

E.1.4. Institute for Cross-Disciplinary Physics and Complex Systems

Dates:(09-11)/07/2019

Responsible: Miguel C. Soriano

Place: Palma de Mallorca, Spain

E.2. Courses

E.2.1. Online Course: Learning From Data (Machine Learning)

Dates:(01/10 - 11/12)/2016

Organized by: California Institute of Technology, USA

Webpage: https://courses.edx.org/certificates/100b5353c990457b912d53a7ca6f6174

E.2.2. Hands-on training workshop: all about eyes

Dates:(01-02)/06/2018

Organized by: IMO

Place: Spain, Barcelona, IMO

E.2.3. Specialization course: Quality in Big Data for Life Sciences

Dates:(11/01 - 10/04)/2019
Organized by: Universitat Autònoma de Barcelona
Webpage: https://e-aules.uab.cat/2018-19/course/view.php?id=15302

E.2.4. Course: PATC: Big Data Analytics

Dates:(05-08)/02/2019
Organized by: Barcelona Supercomputing Center
Place: Spain, Barcelona, UPC Campus Nord
Webpage: https://www.bsc.es/education/training/patc-courses/patc-big-data-analytics-0

E.3. Congresses, meetings, workshops, conferences and Schools

E.3.1. IberSinc Meeting

Dates:(06-07)/10/2016
Organized by: IberSinc
Place: Spain, Tarragona, Escola Tècnica Superior d'Enginyeria
Webpage: http://deim.urv.cat/~alephsys/IBERSINC/reunion.html
Poster contribution: **"Experimental Multistability in Electronic Mackey-Glass Analog Circuit"**

E.3.2. 1st Be-Optical School

Dates:(14-18)/11/2016
Organized by: Be-Optical
Place: Germany, Göttingen, Max Planck Institute for Dynamics and Self-Organization
Webpage: http://beoptical.eu/Public/First_School.html http://www.bmp.ds.mpg.de/beoptical-home.html
Short talk contribution: **"Novel methods for classifying complex data"**

E.3.3. JIPI (Jornada d'Investigadors Predoctorals Interdisciplinària)

Date:9/02/2017
Organized by: Universitat de Barcelona
Place: Spain, Barcelona, historical building of the UB
Webpage: http://www.jipi.cat/en/

E.3.4. 1st BGSMath Data Science Workshop

Date:22/02/2017
Organized by: BGSMath
Place: Spain, Barcelona, Institut d'Estudis Catalans
Webpage: http://bgsmath.cat/1st-bgsmath-data-science/

E.3.5. Opathy 1st Winter School

Dates:(01-04)/03/2017
Organized by: OPATHY ITN
Place: Spain, Barcelona, CRG-Centre for Genomic Regulation
Webpage: https://www.opathy.eu/events/opathy-1st-winter-school-crg-barcelona

E.3.6. DPG Spring Meeting

Dates:(19-24)/03/2017
Organized by: Deutsche Physikalische Gesellschaft (German Physics Association)
Place: Germany, Dresden, TU Dresden
Webpage: http://dresden17.dpg-tagungen.de/

E.3.7. MPIDS retreat 2017

Dates:(03-05)/04/2017
Organized by: Max Planck Institute for Dynamics and Self-Organization
Place: Germany, Werra-Meißner-Kreis, Ludwigstein castle

E.3.8. 2nd Be-Optical School

Dates:(02-05)/05/2017
Organized by: Be-Optical
Place: Poland, Torun, Nicolaus Copernicus University
Webpage: http://beoptical.fizyka.umk.pl/general.html
Talk contribution: **“Unsupervised feature detection in ocular image databases”**

E.3.9. Be-Optical Mid-Term Review

Date:25/09/2017
Organized by: Be-Optical
Place: Spain, Alcúdia, Club Pollentia
Talk contribution: **“Novel unsupervised methods for characterization and classification of ocular images”**

E.3.10. 1st Be-Optical Workshop & Biophysics by the sea

Dates:(25-29)/09/2017
Organized by: Be-Optical & Georg August University
Place: Spain, Alcúdia, Club Pollentia
Webpage: http://beoptical.eu/Public/First_workshop.html
Poster contribution: **“Novel unsupervised methods for characterization and classification of ocular images”**

E.3.11. The 6th International Conference on Complex Networks and Their Applications

Dates:(29/11 - 01/12)/2017

Organized by: Complex Networks

Place: France, Lyon, Université de Lyon 2

Webpage: http://past.complexnetworks.org/index2017.html

Poster contribution: **"Complex Network Analysis of Images of Hunam Retina"**

E.3.12. BIFI International Conference 2018

Dates:(06-08)/02/2018

Organized by: Universidad de Zaragoza

Place: Spain, Zaragoza, Patio de la Infanta

Webpage: http://bifi18.bifi.es/

Poster contribution: **"Novel unsupervised methods for characterization and classification of ocular images"**

E.3.13. AMCOS

Dates:(19-23)/03/2018

Organized by: Marie Curie project COSMOS

Place: Spain, Barcelona, PRBB

Webpage: https://amcosconference.com/

E.3.14. International Conference on BioMedical Photonics

Dates:(16-17)/03/2018

Organized by: Université de Montpellier

Place: France, La grande-Motte, Palais des Congrès

Webpage: http://biomedicalphotonics.org/

Poster contribution: **"Novel unsupervised methods for characterization and classification of ocular images"**

E.3.15. 2nd Be-Optical Workshop

Dates:(18-19)/03/2018

Organized by: Be-Optical

Place: France, La grande-Motte, Hotel Mercure

Webpage: http://beoptical.eu/Public/Second_workshop.html

Talk contribution: **"Novel methods for the characterization and classification of complex images"**

E.3.16. VII Jornada Complexitat.cat

Date:29/05/2018

Organized by: complexitat.cat

Place: Spain, Barcelona, Universitat Pompeu Fabra

Webpage: http://jornada.complexitat.cat/

Poster contribution: **"Complex network analysis of images of human retina"**

E.3.17. Biophysics by the Sea

Dates:(07-12)/10/2018
Organized by: Georg August University
Place: Spain, Alcúdia, Club Pollentia
Webpage: https://www.uni-goettingen.de/en/biophysics+by+the+sea/531969.html
Talk contribution: **"Novel network-based methods for retinal fundus image analysis and classification"**

E.3.18. XIV Jornada de Recerca del Departament de Física

Date:29/01/2019
Organized by: UPC
Place: Spain, Barcelona, Institut d'Estudis Catalans
Poster contribution: **"Complex network analysis of images of human retina"**

E.3.19. 4th BE-OPTICAL Workshop

Dates:(11-12)/03/2019
Organized by: Be-Optical
Place: Germany, Berlin, PicoQuant
Webpage: http://beoptical.eu/Public/Fourth_workshop.html
Talk contribution: **"Outlier mining methods based on network structure analysis"**

E.3.20. 10th International Conference on Complex Networks - COMPLENET'19

Date:21/03/2019
Organized by: CompleNet
Place: Spain, Tarragona, Universitat Rovira i Virgili
Webpage: https://complenet19.weebly.com/
Talk contribution: **"Retinal fundus image analysis and classifcation using complex networks tools"**

E.3.21. JIPI (Jornada d'Investigadors Predoctorals Interdisciplinària)

Date:02/04/2019
Organized by: Universitat de Barcelona
Place: Spain, Barcelona, Universitat de Barcelona
Webpage: https://jipi.cat/en/flashtalks/2019/
Short talk contribution: **"Outlier mining methods based on network structure analysis"**

E.3.22. Complexitat Day 2019

Date:21/05/2019

Organized by: complexitat.cat

Place: Spain, Barcelona, Institut d'Estudis Catalans

Webpage: http://complexitat.cat/jornada/

Poster contribution: **"Outlier mining method based on network structure analysis"**

E.3.23. DATAOPS 2019

Dates:(20-21)/06/2019

Organized by: binlogic

Place: Spain, Barcelona, Moll de Barcelona

Webpage: https://dataops.barcelona/

E.3.24. New trends in biomedical imaging and data analysis Be-Optical Final conference

Dates:(03-04)/07/2019

Organized by: Be-Optical

Place: Germany, Göttingen, Max Planck Institute for Dynamics and Self-Organization

Webpage: http://beoptical.eu/Public/Final_conference/home.html

Poster contribution: **"Outlier mining methods based on network structure analysis"**

E.4. Teaching

E.4.1. Laboratorio de física 1 (Physics laboratory 1)

Semesters:fall 2016, fall 2017 & fall 2018

E.4.2. Laboratorio de física 3 (Physics laboratory 3)

Semester:fall 2019

E.4.3. Laboratorio de ampliación de física (Extension of Physics laboratory)

Semester:fall 2019