• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

MARIANA ANTUNES NOGUEIRA

Creating Evaluation Functions for Oncological Diseases based on PET/CT

Dissertação apresentada à Universidade de Coimbra
para cumprimento dos requisitos necessários à obtenção
do grau de Mestre em Engenharia Biomédica

Thesis submitted to the University of Coimbra for
compliance with the requirements for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. **Pedro Henriques Abreu**
Prof. Dr. **Pedro Martins**

Coimbra, 2015

Este trabalho foi desenvolvido em colaboração com:
This work has been developed in cooperation with:

Centre for Informatics and Systems of the University of Coimbra

Abstract

Oncological diseases remain a worldwide leading cause of death, despite advances in treatment techniques. In this context, treatment response is naturally a very attractive research topic – the ultimate goal is set on finding, within patient and pathology characteristics, good predictors of treatment response, so as to help deciding on the best treatment approach for each patient.

PET/CT imaging is the basis for the evaluation of response-to-treatment of several oncological diseases. In practice, such evaluation is manually performed by specialists, which is rather complex and time-consuming. As alternatives, evaluation measures of lesion malignancy – such as the popular SUV – have been proposed, but present questionable reliability – in the case of SUV, this is due to multiple sources of variability.

The aim of this project is to walk towards a reliable evaluation function of treatment response, based on the before and after-treatment values of a set of clinical variables and image features of the lesions (extracted from PET/CT images), and using evolutionary approaches for regression. Clinical data used in this project was provided by IPO-Porto, comprising a total of 63 patients and 2 distinct oncological pathologies – Hodgkyn lymphoma and neuroendocrine tumors.

The preliminary results concerning the proposed approach are optimistic – an evaluation function with class-wise accuracies of 80%, 75%, 85,71% and 88,89% (for a problem with 4 treatment response classes), was obtained.

Resumo

As doenças oncológicas mantêm-se no *top* das maiores causas de morte a nível mundial, apesar dos avanços a nível de técnicas de tratamento. Neste contexto, a resposta a tratamento é um tópico de estudo muito atractivo – o objectivo último está fixado em encontrar, entre características do paciente e da patologia, bons preditores da resposta a tratamento, para auxiliar na decisão sobre a melhor estratégia de tratamento para cada paciente.

A PET/CT é base para a avaliação da resposta a tratamento de uma grande quantitdade de doenças oncológicas. Na prática, essa avaliação é efectuada manualmente por especialistas, o que consiste numa tarefa complexa e morosa. Como alternativas, medidas de avaliação da malignidade das lesões – como o SUV – foram propostas, mas apresentam uma confiança questionável – no caso do SUV, isto deve-se a múltiplas fontes de variabilidade.

O objectivo deste projecto é caminhar na direcção de uma função de avaliação de resposta a tratamento de confiança, baseada nos valores de antes e depois de tratamento de uma série de variáveis clínicas e *features* de imagem das lesões (extraídas de PET/CT), e adoptando abordagens evolucionárias para regressão. Os dados clínicos usados neste projecto foram fornecidos pelo IPO-Porto, compreendendo 63 pacientes e 2 patologias oncológicas – linfoma de Hodgkyn e tumores neuroendócrinos.

Os resultados preliminares são optimistas – uma função de avaliação com *accuracies* por classe de 80%, 75%, 85,71% and 88,89% (para um problema com 4 classes de resposta a tratamento), foi obtida.

Acknowledgements

I would like to express my gratitude to Professors Pedro Henriques Abreu and Pedro Martins for the unconditional availability and support. I would also like to acknowledge Professor Penousal Machado for always helping when necessary.

I must thank the medical team of IPO-Porto, with a special mention for Doctor Hugo Duarte, for the availability and cooperation.

Finally, I would like to thank my family and friends, who make it all worth it.

Contents

Acronyms

AD Alzheimer's Disease.

ANN Artificial Neural Networks.

AUC Area Under the Curve.

BCDR Breast Cancer Digital Repository.

CAD Computer-aided Diagnosis.

CM Cerebral Microangiopathies.

CN Cognitive Normal.

DCT Discrete Cosine Transform.

DDSM Digital Database for Screening Mammography.

DSS Dice Similarity Score.

FDG Fludeoxyglucose (18F).

FPR False Positive Rate.

GA Genetic Algorithm.

GATM Generic Algorithm Template Matching.

GLCM Gray-level Co-occurrence Matrix.

GLRL Gray-level Run-length Matrix.

HGD Histograms of Gradient Divergence.

HOG Histogram of Oriented Gradients.

ICA Independent Component Analysis.

IRMA Image Retrieval in Medical Applications.

kNN k Nearest Neighbors.

LBP Local Binary Patterns.

LDA Linear Discriminant Analysis.

LVQNN Learning Vector Quantization Neural Network.

MCI Mild Cognitive Impairment.

MIAS Mammographic Image Analysis Society.

MIM Mutual Information Maximization.

MLP Multilayer Perceptron.

MR Magnetic Resonance.

MRI Magnetic Resonance Images.

mRMR Minimal Redundancy Maximal Relevance.

MS Multiple Sclerosis.

NEAT NeuroEvolution of Augmenting Topologies.

PBCC Point Biserial Correlation Coefficient.

PCA Principal Component Analysis.

PET/CT Positron Emission Tomography – Computed Tomography.

PNN Probabilistic Neural Network.

RBFNN Radial Basis Function Neural Network.

ROC Receiver Operating Characteristic.

ROI Region of Interest.

SMOTE Synthetic Minority Over-sampling Technique.

SUV Standardized Uptake Value.

SVM Support Vector Machines.

TNR True Negative Rate.

TPR True Positive Rate.

WMH White Matter Hyperintensities.

List of Figures

List of Tables

Chapter 1

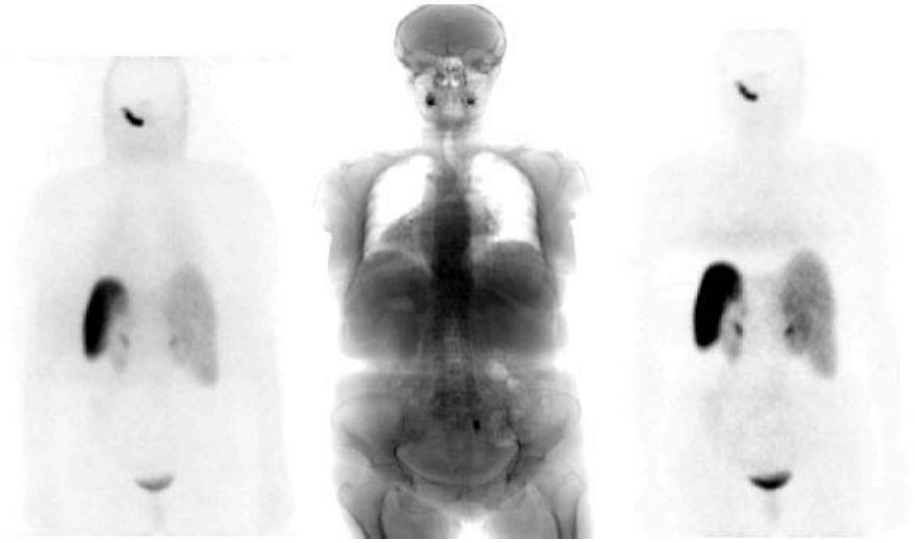
Introduction

Oncological diseases remain a worldwide leading cause of death, despite advances in treatment techniques [1]. As of 2012, they were responsible for 14.6% of all human deaths, and about 14.1 million new cases occurred at a global scale [2]. Incidence rates are increasing, as more people live to an old age and as lifestyle changes occur in the developing world [3]. In terms of financial burden, the costs of cancer have been estimated at over €1 trillion per year as of 2010 [4]. In this context, treatment response is naturally a very attractive research topic – the ultimate goal is set on finding predictive models of treatment response based on patient and pathology features, in order to personalize treatment strategies aiming at the best possible outcome for the patient.

1.1 Contextualization

Positron Emission Tomography – Computed Tomography (PET/CT) imaging is the basis for the diagnosis and staging of several oncological diseases. In PET, one takes advantage of prior knowledge regarding functional properties of pathologies, namely the abnormal uptake of certain substances by some tissues, and uses radioactively-labelled analogues of those substances to visualize their distributions throughout the organism. In this project, two oncological diseases are used as case studies: Hodgkyn lymphoma and neuroendocrine tumors. For both cases, all clinical data was provided by IPO-Porto. For a better contextualization, a more detailded descripition of these two pathologies is given.

Hodgkin lymphoma has its origin in the lymphatic system. Lymphoid tissue can be found in several organs, such as tonsils, thymus and spleen, as well as in lymph nodes all over the body. Thus, in theory, Hodgkin disease can start almost anywhere in the body. The most usual is starting in lymph nodes and spreading to other lymph nodes through lymphatic vessels which connect them. Hodgkin lymphoma lesions are characterized by very high metabolic activities. For that reason, the classical tracer used in PET is Fludeoxyglucose (18F) (FDG), a glucose analogue which is retained by tissues of high metabolic activity. It is important to be aware of the normal distribution of the tracer, i.e., in healthy patients, so as to differentiate physiological from pathological uptakes – some tissues such as bladder, lung, heart and brain tissue naturally uptake great glucose quantities.



(a) PET (b) CT (c) PET/CT

Figure 1.1: Examples of PET,CT and PET/CT coronal slices of a patient suffering from paraganglioma, a neuroendocrine tumor that affects head and neck (head in this particular case).

Neurondocrine tumors start in the endocrine and nervous systems. They are mostly found in the pancreas, intestine and lungs, but can occur everywhere in the endocrine system. Neuroendocrine tumors are associated with the overexpression of somatostine receptors. For that reason, Gallium-68, an analogue of somatostine, was selected as tracer for PET. Once again, one should be aware that there are regions of naturally high uptake (e.g. in exocrine pancreas or adrenal glands) in order to be able to distinguish physiological from pathological uptake.

The precise localization of the anomalies is difficult with only PET information. CT, on the other hand, is an anatomical modality. It takes advantage of the relative x-ray absorption by tissues to produce a density image - denser regions such as bone absorb more x-rays and thus present higher intensities, and less dense regions such as lungs (due to air percentage) appear as dark regions. PET/CT fusion is becoming a popular practice, as it couples the benefits of the two modalities - functional analysis with precise anatomical location. Examples of PET,CT and PET/CT coronal slices of a patient suffering from paraganglioma, a neuroendocrine tumor that affects head and neck (head in this particular case) are presented in Figures 1.1a – 1.1c.

In addition to diagnosis and staging, PET/CT is becoming important in the evaluation of response to treatment – changes in tracer uptake by tumors have been shown to be useful for assessing response to therapy. In practice, lesions are evaluated via visual inspection by specialists, which is rather time-consuming. As tracer uptake is related to the malignancy of lesions, some have proposed quantitative-driven measures of tracer uptake as a potential alternative for faster assessment of lesion malignancy. The exact radioactivity concentration within the lesion, provided by the PET scanner, would not be robust to patient size and amount of injected tracer, i.e., for two patients with similar body tracer distributions:

1. if the amount of injected tracer is the same for both, activity concentration would be

higher for a smaller-sized patient.

2. for same-size patients, activity concentration will be higher for a patient who was injected a higher amount of tracer.

In fact, patient size and the amount of injected tracer are the two main sources of concentration variation. To compensate for such variations, a normalization of the activity concentration with respect to patient size and amount of injected tracer was proposed. Such normalized measure is known as Standardized Uptake Value (SUV), and has been widely adopted.

In the original formulation of SUV, patient weight is used as the patient size normalization factor. However, since tracer is mainly uptaken by non-fat mass, lesion SUV is likely overestimated in patients with higher body fat percentage. For that reason, some alternatives to patient weight, namely lean body mass and body surface area have been suggested.

Theoretically, SUV is expected to provide a good evaluation of lesion malignancy, and that is why it has been used in many studies with related purposes. However, there are other (not so obvious, but still relevant) possible sources of SUV variability, aside from patient size and amount of injected tracer [5]. These can be related to:

1. Imaging Physics – e.g. limited detector resolution leads to partial volume effects, i.e., the measured SUV is increasingly reduced for objects of decreasing size;

2. Patient Status – e.g. blood glucose levels as well as endogenous insulin levels can result in lesion-SUV underestimation, as in such cases glucose is preferably stored in muscle and less available for other tissues such as tumor tissue;

3. Scan Protocol – errors in several components of the scan protocol later reflect on the measured lesion SUV, such as: 1) time interval between injection and image acquisition; 2) measurement of the residual activity in the syringe; 3) measurement of patient weight; 4) synchronization of clocks used for dose assays and scanning; 5) patient respiratory motion; 6) data entry;

4. Data Processing – errors can arise in the two main steps of the data processing stage: 1) corrections for confounding effects like attenuation, scattered and random coincidences, scanner deadtime, and detector efficiency variations; 2) image reconstruction, where the raw scanner data is converted into standard units that are typically related to the scanner, the reconstruction method and the reconstruction parameters;

5. Scanner Calibration – in order to obtain the image in SUV units, two steps are necessary: 1) estimation of the scanner calibration factor, so as to convert from scanner units to radioactivity concentration units; 2) conversion from radioactivity concentration units to SUV units; An erroneous calibration factor can be reflected in the measured SUV;

6. Analysis Methods – noise and limited resolution often make it difficult to draw a boundary of the region of interest with certainty. On the other hand, average SUV can be very dependent on the defined boundary. For that reason, it has become more common to take the maximum SUV within boundaries as the lesion SUV, as this one is invariant to small boundary shifts. However, some concerns arise from the fact that we are basing the evaluation of our lesion in a sole pixel, regarding robustness to noise.

Some of these errors have effects of only about 5% in the measured SUV, but others range up to 50% or more. In such context, the need for searching for more reliable metrics of lesion evaluation arises.

1.2 Objectives

In the lack of a reliable evaluation metric, the main objective of this project is to collect relevant features and set up an accurate evaluation function for the automatic evaluation of tumor treatment response. This main objective can be divided into a set of sub-objectives:

1. Identification of the candidate features, which comprehends:

    a) Data collection – clinical data regarding 63 patients suffering from oncological pathologies was collected and provided to us by a nuclear medicine team of IPO-Porto. With the help of the medical team, a set of clinical variables were selected to be incorporated in this study. Also, PET/CT images were provided.

    b) Feature extraction – in addition to the clinical variables defined by the medical team, image features of the lesions were extracted from the PET/CT.

    c) Eventual dataset transformations (e.g. dimensionality reduction).

2. Validation of the feature set – before the application of evolutionary approaches for obtaining the evaluation function, we aim to prove that our pool of features contains good predictors for the treatment response class, through a classification experiment.

3. Obtaining the evaluation function – after the validation of our features, we aim to obtain the evaluation function, adopting an evolutionary approach for regression.

1.3 Research Contributions

The work in this project has resulted in the development of three papers:

- Mariana A. Nogueira, Pedro Henriques Abreu, Pedro Martins, Penousal Machado, Hugo Duarte, João Santos. An Artificial Neural Networks Approach for Automatic Assessment of Treatment Response in Oncological Patients using PET/CT Images, Special issue on Learning From Medical Imaging, Neurocomputing (IF: 2.083) (Accepted with major revision)

- Mariana A. Nogueira, Pedro Henriques Abreu, Pedro Martins, Penousal Machado, Hugo Duarte, João Santos. Image Descriptors in Healthcare Contexts: A Systematic Review, Journal of Biomedical Informatics (IF: 2.194) (Submitted on 26/4/2015)

- Mariana A. Nogueira, Pedro Henriques Abreu, Pedro Martins, Penousal Machado, Hugo Duarte, João Santos. Creating Evaluation Functions for Oncological Diseases based on PET/CT (To be Submitted)

1.4 Document Structure

The remainder of the document is organized as follows: in Chapter 2, some useful background knowledge for a better understanding of the remainder of the document will be provided; in Chapter 3, a literature review regarding the application of image descriptors in the field of medical image analysis will be performed; in chapter 4, the adopted methodology and parameterization for the data processing, classification and regression stages is reported; in chapter 5, the results of the experiments are reported and discussed. Finally, chapter 6 is dedicated to the main conclusions and future work.

Chapter 2

Background knowledge

In this chapter, some useful background knowledge is provided for a better understanding of the remainder of the document.

2.1 Image Segmentation

Image segmentation corresponds to the act of extracting a region of interest out of an entire image. In our project, the aim is to segment lesion regions, for further extraction of descriptive image features of the lesions. A quality segmentation is of major importance, as the values of the extracted features are very dependent on the segmented region, and this can seriously compromise their discriminative ability and consequently the classification results. The quality of manual segmentation by specialists has not yet been matched by any proposed alternative; however we did not have access to annotated images. Thus, we opted for an automatic segmentation algorithm. Segmentation was not the scope of this project, and, as such, we decided to adopt a classical automated segmentation algorithm, the region-growing algorithm. It basically consists of growing a list of pixel locations from a single intensity maximum – seed point – by appending immediate neighboring pixels with intensity levels above a certain threshold – which is a percentage of the maximum –, doing the same for the newly added pixels, and so on.

2.2 Image Descriptors

The selection of the appropriate image descriptors is a rather decisive stage in our project. In this section, we provide some theoretical background on the image descriptors selected for our project, as well as some examples of features that are usually extracted from their outputs. For reasons related to characteristics of our dataset (pointed out in 4.1), shape features could not be used. We selected 6 texture descriptors with solid reputations in medical image analysis applications, as can be confirmed in the literature review (chapter 3). These are:

Gray-level Histogram The gray-level histogram is a vector containing the absolute frequency of each gray level in the segmented patch. A few descriptive features of the gray-level distribution, such as mean, standard deviation, skewness and kurtosis are typically computed.

Gray-level Co-occurrence Matrix (GLCM) Given a patch with gray-levels in the range $[a, b]$, the GLCM [6] is an $a \times b$ matrix containing, in the element (i, j) – with $i, j \in [a, b]$ – the number of times a pixel of intensity i is at distance d from a pixel of intensity j, in a pre-defined direction θ. Typically, a set of 14 texture features proposed by Haralick [6] – known as Haralick features – is extracted from the GLCM. To achieve some kind of rotation invariance, features are usually computed for different direction GLCMs (for instance 0^{o}, 45^{o}, 90^{o} and 135^{o}) and averaged over all directions.

Gray-level Run-length Matrix (GLRL) [7] A gray-level run is a sequence of consecutive pixels with the same gray-level, in a certain direction. The GLRL is a matrix that contains, in the element (i, j), the number of j-length runs of pixels with the gray-level i, in a pre-defined direction θ. A set of run length features proposed by Galloway [7] is usually extracted from this matrix. Usually, features are computed for different directions and averaged over them, with the purpose of achieving rotation invariance.

Wavelets 2D discrete wavelet decomposition consists of two successive 1D wavelet decompositions, one in the horizontal direction and the other in the vertical direction of the patch matrix. First, each row of the matrix goes through a lowpass and a highpass filter, and their outputs are subsampled by a factor of 2 (as half the samples are sufficient to reconstruct the signal, according to Nyquist's theorem). Hence, each row will yield two vectors of half its size. The ones that underwent lowpass filtering contain the low frequency information of the corresponding row and thus represent a coarse approximation to the row itself. On the other hand, the vectors that underwent highpass filtering contain the high frequency information, such as edges. Assembling the lowpass vectors, one obtains a coarse approximation of the initial patch matrix, with the same number of rows but only half the columns. Assembling the highpass vectors, one gets a matrix with the high frequency components of the initial image, such as edges, also with the same number of rows and half the columns of the original one. Then, the columns of these matrices are decomposed as well. Following the same logic, each of the two matrices will generate two matrices with half its number of rows. We have then 4 matrices with half the resolution of the initial image matrix, each one with the following characteristics:

1. The one which went through lowpass filtering in both directions is the approximation matrix, and consists of a coarse approximation of the original image, with half the resolution;

2. The one which went through lowpass filtering in the horizontal direction and highpass filtering in the vertical direction consists of the vertical detail matrix;

3. The one which went through highpass filtering in the horizontal direction and lowpass filtering in the vertical direction consists of the horizontal detail matrix;

4. The one which went through highpass filtering on both directions is the diagonal detail matrix.

This is the result of a single level decomposition. Decompositions to further levels are obtained by successively applying the same mechanism to the previous level's approximation matrix.

A wavelet family must be chosen – it defines the morphology of the lowpass and highpass filters. The most widely used in biomedical signal/image analysis applications are Haar and Daubechies families. Usually, features as mean, standard deviation, energy and entropy are extracted from each matrix at the different levels.

Gabor filters Gabor filters are linear filters which consist of Gaussian kernels modulated by a sinusoidal plane wave, and are very popular in edge detection. Usually, several filters with different frequencies and orientations are generated and convolved with the image of interest, and features such as mean, standard deviation, energy and entropy are computed from the response images.

Local Binary Patterns (LBP) The LBP [8] operator is characterized by a radius R and a neighborhood P, which are correlated: the operator is centered on a pixel; if the radius is 1, the 8 immediate neighbors of the central pixel will be considered. If the radius is 2, the 16 pixels with one pixel of interval from the central pixel will be considered, and so on.

Usually, the patch is divided into blocks before LBP computation. Given a block, the center of the LBP operator will travel through all pixels and, for each pixel, a binary string of length P will be produced. Each of the P neighbors is responsible for a bit: 0 if the neighbor's gray level is inferior to the central pixel's, and 1 otherwise. This binary string is usually converted to its decimal form.

Then, for each block, a histogram is produced, with the count of the number of times each decimal number is produced by pixels in the block.

This descriptor is intensity- and scale-invariant. In addition, it can also be rotation invariant, if we only consider the so-called uniform patterns. These consist of patterns with no more than 2 bit transitions in the binary number, and are the most occurring patterns – for R=8, there are 58. A histogram is usually produced with the count of each of the 58 uniform patterns, and 1 histogram slot is reserved for the remaining patterns.

2.3 Dimensionality Reduction

The complexity of a classification process is said to exponentially increase with the problem's dimensionality. As such, and since high-dimensional datasets often present some redundant and even irrelevant information, some methods exist with the aim of reducing dimensionality by eliminating such information and preserving only the most relevant. In this project, we selected a classic method, Principal Component Analysis (PCA). In PCA, data is projected onto a new axes system, given by the directions of highest data variance, with the hypothesis that, since data are more spread along these directions, they will facilitate class separation. Such axes system consists of the eigenvectors of the data covariance matrix, with the eigenvalue magnitude being a measure of the data variance encompassed in the direction of the corresponding eigenvector – the eigenvector associated with the highest-magnitude eigenvalue corresponds to the direction of highest data variance, and so on. The goal is to discard directions of little contribution to overall data variance, reducing dimensionality but at the same time preserving most of data variance. By projecting data onto these directions, one loses the concept of features, and each dimension represents a component.

Several methods have been proposed for selecting the final number of components. One of the most popular is the Scree plot, where one plots eigenvalues' magnitudes versus the

corresponding components and observes at what point it goes off, i.e., it drops down to a level that can be visually interpreted as zero. Another popular criterion is assuring that we preserve at least 95% of data variance.

2.4 Synthetic Data Generation

Data imbalancement, i.e., not having a balanced number of instances among classes, may result in poor training regarding minority classes, and in a consequent tendency for bad generalization with respect to minority-class samples. In order to minimize this effect, one can oversample the minority class, undersample the majority class, or use more sophisticated methods and generate synthetic data.

We decided to generate synthetic data – we chose a very popular technique, the Synthetic Minority Over-sampling Technique (SMOTE) [9]. This algorithm generates a synthetic sample of a class by randomly picking two samples of that class and adding to one of them a percentage of the difference between the two. Graphically, the synthetic sample corresponds to a random point in a line segment connecting the two randomly chosen samples. This method allows us to obtain smoother decision regions than if simple over-sampling was applied.

2.5 Classification

Classifiers are a class of algorithms which can be trained by examples in order to learn to classify objects based on a set of features. Literature is vast on this subject: Fisher's Linear Discriminant, Naive Bayes, k Nearest Neighbors (kNN), Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), are just some examples of prominent classifiers. The goal of the classification stage in the context of our project is to validate our set of features – before the application of evolutionary approaches for obtaining the evaluation function, we aim to evaluate if our pool of features contains good predictors for the treatment response class. For that reason we used a few standard architectures of the complex, but with solid reputation, ANNs, and a baseline classifier, kNN.

2.5.1 Artificial Neural Networks

A generic ANN is composed of one or more layers of neurons. Feature vectors are fed to the first layer. Each feature of the vector is assigned a weight by each neuron of this layer, which outputs a value based on the response of an activation function to the weighted sum of all features (plus an optional bias). The selection of such activation function is dependent on the specific problem we are facing. These outputs can already be the final outputs of the network, or can be fed to another layer and be used as inputs to produce new outputs through the same mechanism. Finally, the outputs of the last layer are the network outputs. This layer is the output layer and other layers are referred to as hidden layers.

In classification tasks, one wants the ANN to learn to correctly classify an object based on a set of its features. For such learning to be possible, a training process must occur, in which ANN are fed input-output examples and the weights and bias of the neurons are iteratively updated in the sense of minimizing the error between the obtained and expected output.

We selected four standard neural network architectures: the Multilayer Perceptron (MLP), Learning Vector Quantization Neural Network (LVQNN), Radial Basis Function Neural Network (RBFNN) and Probabilistic Neural Network (PNN).

The perceptron is the simplest neural network architecture, consisting of only one neuron layer. It can only solve linearly separable problems. By adding neuron layers to the perceptron, the network starts being able to solve non-linearly separable problems – the MLP is usually applied in scenarios where data is not linearly separable and, normally, the higher the degree of data nonlinearity, the higher the number of necessary neurons and layers for data separation. However, in many scenarios one-hidden-layer MLP with only a few neurons in the hidden layer can achieve very high performances. The MLP is typically trained with a backpropagation algorithm.

LVQNN, RBFNN and PNN, on the other hand, are one-hidden-layer networks with distance-based training methods. These architectures are usually adopted in situations where data is distributed in such way that we can divide it in several clusters – more clusters than classes and non-linearly separable class-wise. In such context, we can see neurons from the hidden layer as points in the feature space (the coordinates are the weights). Ideally, there will be a neuron close to each cluster in order to be activated when presented to an input vector closer to that cluster than to the others. If possible, one should *a priori* inspect the most appropriate number of clusters to divide the data, since this is the number of neurons that should be used in the hidden layer of these networks. The placement of the neurons near the appropriate clusters is performed in different ways in different networks:

1. In LVQNN, a learning rule exists that, based on training samples, updates the weights of the neurons in order to get closer or farther from them depending on whether the desired output was produced or not. Training set imbalancement can highly affect LVQNN, as minority-class training samples will not suffice to update weights properly;

2. In RBF-based neurons, clustering techniques such as k-means or subtractive clustering are usually adopted for finding the optimal placement of the neurons. Then, a spread value – which corresponds to the radius of the Gaussian kernel of a RBF neuron – is associated with those neurons. An input vector will activate a neuron if it is under its radius. That activation will be stronger the closer the input vector is to the center of the neuron. Thus, the spread value should not be too small, in order not to activate neurons only at their centers (overfitting), neither too big, for preventing the simultaneous activation of neurons that should respond in different ways to an input vector;

3. The PNN has a RBF-neuron centered in each training sample. Based on the distance of the test sample to all training samples and their labels, the RBF layer will output a vector of class probabilities, and a competitive layer will select the highest probability class. PNN is usually more accurate than RBFNN, however it is impractical for large training sets.

2.5.2 k Nearest Neighbors

The kNN is a distance-based majority-voting classification algorithm, composed of the following steps:

1. The training set is stored, i.e., all training samples and corresponding label;

2. Given a test sample, the distances between such sample and all training samples are computed. This step is limiting for the application of kNN in the presence of large datasets;

3. Finally, the test sample is assigned the most represented class in the k nearest training samples.

2.6 Genetic Algorithms

In the final stage of our project, we will use symbolic regression for obtaining the evaluation function. Symbolic regression is a particular derivation of Genetic Algorithm (GA). In this section, we provide some basic knowledge on GAs in general, and symbolic regression in particular.

2.6.1 Introduction to Genetic Algorithms

GAs are optimization algorithms inspired in the natural selection theory of evolution. According to this theory, evolution towards fitter generations can be explained by the survival and reproduction of the fittest individuals.

In an optimization process, we normally identify a set of parameters which can be tuned in order to optimize results. A candidate solution is any set of values for those parameters. In GAs, candidate solutions are encoded in bit-strings, consisting of concatenations of parameter values in binary format. Such a bit-string is called a chromossome (each bit representing a gene), and represents an individual. A GA usually starts with a pre-defined number of randomly initialized individuals – the initial population. The latter must be large and diverse enough to allow evolution towards fitter individuals. The fitness of each individual is assessed through a fitness function – a function which somehow measures how close the output is to the desired result. Then, chromossome selection for reproduction takes place – it is imposed that fitter individuals are selected with much higher probability. Three main genetic operators can actuate over each selected parent, so as to generate the next-generation population – copy, crossover and mutation– with pre-defined probabilities (mutation is normally a rare process). Copy refers to the simple placement of an identical individual in the next-generation population; crossover is the mutual exchange of genes between two chromossomes, creating two different chromossomes; mutation is the punctual alteration of a bit value in the bit-string of the chromossome. Until the optimal solution is found or a stopping condition is verified (e.g. a maximum number of generations), the population will evolve through this mechanism towards future generations.

2.6.2 Symbolic Regression

Symbolic regression is a particular application of GAs, where the goal is to evolve mathematical expressions that best fit the relationship between a set of inputs and a set of outputs, i.e., regression. In this type of problems, two fundamental elements are required:

1. The terminal set – consists of the set of the input features and some constants. String representations of the features are encoded in the chromossomes (e.g. $X1$, $X2$, ..., Xn, for a dataset with n features).

2. The function set – consists of the operators that can be selected for the mathematical expression (e.g. $+, -, /, *$). In our experiments, the function set consisted of $+, -, *, /$, natural logarithm, sine, cosine, sigmoid function and if-then-else statement.

The chromossomes are composed of elements from the terminal and the function set, thus mathematical expressions that are functions of the features. A tree representation is usually adopted for individuals. The outputs of such expressions are computed for every input-output example and later compared with the expected outputs. The fitness function is usually the comparison measure, for instance an error measure between obtained and real outputs, that we want to minimize. In this project, the error measure the sum of absolute differences between obtained and real outputs.

2.7 Sampling Strategies

Cross-validation is the most common form of model validation – the dataset is divided in two sets – one set is used for training the model (training set) and the other is used for testing (test set). The fact that the test set is independent from the training set can be a good indicator for the generalization ability of the model. As the results are always in some way dependent on the training and testing sets, i.e., different results may outcome for different dataset partitions, one usually assesses the model's performance for several different dataset partitions, and takes average performance as the true model performance. Probably the most popular form of cross-validation is the so-called k-fold cross-validation: in k-fold cross-validation data is divided into k subsets (folds), and each is fold is once used as the test set while the remaining folds are used for training the model. Leave-One-Out cross-validation is a special case of k-fold cross-validation, for $k = N$, i.e., the total number of available examples. In this project, the Leave-One-Out approach was used, for reasons related with the characteristics of our dataset, pointed out in 4.1.

2.8 Performance Measures

The performance evaluation of a classifier is normally based on a confusion matrix (Figure 2.1). This matrix the distribution of actual vs. predicted classes of the test samples.

Based on that matrix, several useful evaluation metrics can be derived. Precision shows the proportion of correctly predicted positive cases relative to all the predicted positive ones (equation 2.1).

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

True Positive Rate (TPR) or Sensitivity represents how many positive examples the classifier was able to correctly identify (equation 2.2).

$$TPR = \frac{TP}{TP + FN} \tag{2.2}$$

True Negative Rate (TNR)1 or Specificity represents how accurately the classifier behaves in terms of predicting the negative class (equation 6).

$$TNR = \frac{TN}{TN + FP} \tag{2.3}$$

False Positive Rate (FPR) corresponds to the percentage of misclassified samples of the negative class (equation 2.4).

$$FPR = \frac{FP}{FP + TN} \tag{2.4}$$

Accuracy represents how many predictions of the classifier were in fact correct (equation 2.5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

Table 2.1: Typical confusion matrix.

		Actual Class	
		Negative	Positive
Predicted Class	Negative	True Negative (TN)	False negative (FN)
	Positive	True positive (TP)	False positive (FP)

The overall accuracy shall not be a good performance measure of the classifier in imbalanced data scenarios, as one can achieve high overall accuracies by classifying all samples onto majority-class, which is not desired. In such scenarios, the F-measure can be adopted, as it provides a balance between precision and sensitivity (equation 2.6).

$$F - measure = \frac{2 * precision * sensitiviy}{precision + sensitivity} \tag{2.6}$$

In this specific problem, we have 4 classes. Except for accuracy, all other enumerated metrics are defined for 2-class problems – at least in terms of overall performance evaluation, i.e., they could be computed to assess the performance of the algorithm towards each class. Given that our dataset is highly imbalanced, the overall accuracy shall not be a good performance measure. We decided to compute the 4 class-wise accuracies and base comparison on measures that take into account a balance among them.

2.9 Classifier Comparison

In this project, classifier comparison will be fundamentally based on Friedman's non-parametric test, followed by the Bonferri-Dunn test [10].

Friedman's non-parametric test for comparison of k classifiers, based on their performances in N experiences, can be summarized in the following steps:

1. The k classifiers are ranked for each of the N experiments, based on their scores. In the presence of ties, the involved ranks are averaged.

2. The Friedman statistic ($T1$) is computed based on such ranks, and is compared with the F distribution for $k-1$ and $(k-1)*(N-1)$ degrees of freedom, so as to infer whether there is or not a significant statistical difference among the scores of the classifiers – one can reject the null hypothesis that there is no significant statistical difference among the performances of the k classifiers if the $T1$ is superior to $F(k-1,(k-1)*(N-1))$.

Then, if the null hypothesis is to be rejected, a Bonferri-Dunn test can be performed so as to determine which classifier(s) significantly outperform a baseline classifier (under a predefined significance level) – a critical value (CD) for the difference of the average ranks between the baseline and other classifiers is computed. Only those classifiers whose average rank is better to that of the baseline classifier by more than the critical value are considered to significantly outperform it.

Chapter 3

Literature Review

In this project, image features are based on image descriptors. The latter are algorithms that process images and output information which can be used to extract descriptive features of intensity, texture and shape, of objects in the image, for example. Given these properties, image descriptors can be very useful in medical image analysis. In fact, they have already been applied in several medical image analysis studies, for instance in the analysis of breast lesions in mammographic images, oncological diseases in PET, lung pathologies in CT, and brain pathologies in MR images. In this chapter, a brief literature review is performed, featuring several studies that are illustrative of current applications of image descriptors in medical image analysis, with the objective of finding the most suited descriptors for our problem. For an easier reading, the text is organized in different sections regarding different imaging modalities – mammography, PET, CT and Magnetic Resonance Images (MRI). We decided to narrow down our choice to these image exams, as they are illustrative of the diversity and relevance of works based on the use of image descriptors. Towards the end of the chapter, two sections are reserved specifically for some examples of the application of ANN and GAs, in medical image analysis.

On the specific problem of tumor treatment response analysis, little documentation exists. In the PET section, we report an example, by Naqa et al. [21], where logistic regression is used for obtaining a predictive model of treatment outcome based on pre-treatment image features. In our project, before and after-treatment image features will be extracted, and an evolutionary approach for regression (symbolic regression) will be performed for obtaining an evaluation function of treatment response.

3.1 Mammographic Images

Mammographic images are generally used for the analysis of lesions related to breast cancer, which usually correspond to masses and/or calcifications. These lesions stand out as more intense regions due to their higher density (and thus higher x-ray absorption), when compared to normal breast tissue. For this section, we selected three representative studies, whose main features are summarized in Table 3.1.

Moura and Guevara-Lopez [11] performed a comparison of the suitability of a comprehensive set of image descriptors for the detection of breast lesions in general, and that of masses or calcifications in particular. Moreover, they explored the hypothesis that combining other forms of clinical data with the information extracted from image descriptors

Table 3.1: Main features of the studies reviewed in section 3.1.

Authors	Objectives	Dataset(s)	Descriptors Families
Moura & López [11]	A comparison of the suitability of a very comprehensive set of image descriptors for detecting lesions in general and masses or calcifications in particular; improving performance by combining other clinical information with the image descriptors.	913 segmentations of benign findings and 849 of malignant findings, extracted from the digital database for screening mammography (DDSM) [12]. 187 benign findings and 175 malignant findings from the BCDR-F01 dataset of the breast cancer digital repository (BCDR) [13]. Both databases are public repositories.	Gray-level matrix Gray-level histogram Invariant moments Zernike moments GLCM GLDM GLRL Gabor filters Wavelets Curvelets HOG HGD (novel)
Tahmasbi, Saki & Shokouhi [14]	Extraction of shape and margin properties of breast lesions using Zernike moments, for discriminating between benign and malignant lesions.	209 normal breasts, 67 regions of interest (ROIs) with benign lesions and 54 ROIs with malignant lesions, extracted from digital mammograms from the Mammographic Image Analysis Society (MIAS) database [15].	Zernike moments
Sharma & Khanna [16]	Zernike moments for malignancy classification of mammographic patches.	534 malignant and 266 benign samples from IRMA database. 407 benign and 857 malignant findings from DDSM database.	Zernike moments

can lead to improved performances. The patches of the lesions were manually segmented from mammograms extracted from Digital Database for Screening Mammography (DDSM) DDSM [12] and Breast Cancer Digital Repository (BCDR) [13]. Then, 12 image descriptors were computed from the segmented patches: gray-level vector of intensities (IS), gray-level histogram, Hu's invariant moments [17], Zernike moments [18], GLCM, gray-level difference matrix, GLRL, Gabor filter banks, wavelets, curvelets, Histogram of Oriented Gradients (HOG) [19], and a novel descriptor named Histograms of Gradient Divergence (HGD). The final features for each descriptor were sets of statistics computed from their outputs. For classification, SVM, Random Forests, Logistic Model Trees, kNN and Naive Bayes classifiers were used, and the highest Area Under the Curve (AUC) among them was the final performance measure for each feature vector. The main results are presented in Tables 3.2-3.4.

Table 3.2: Performance of standalone clinical data for each scenario.

Dataset	All lesions	Masses	Calcifications
DDSM	0.853	0.867	0.807
BCDR	0.712	0.829	0.725

Table 3.3: Group of standalone descriptors that significantly outperform the remainder, for each scenario. Within each scenario, the highest performing standalone descriptor(s) is(are) presented first, with the corresponding performance between brackets. Then, the names of other descriptors with similar performances are enumerated.

Dataset	All lesions	Masses	Calcifications
DDSM	GLRL (0.743)	GLRL (0.733)	Wav. (0.733)
	HGD	HGD	GLCM
BCDR	HGD (0.825)	HGD (0.860)	GLCM (0.793)
	HOG	HOG	Gabor, HGD, Wav., Curv.

Table 3.4: Group of *clinical data + descriptor* combinations that significantly outperform the remainder, for each scenario. Within each scenario, the highest performing combination(s) is(are) presented first, with the corresponding performance between brackets. Then, the names of other combinations with similar performances are enumerated.

Dataset	All lesions	Masses	Calcifications
DDSM	IS (0.868)	IS, Zer (0.89)	Gab (0.803)
		HGD	
BCDR	HGD (0.817)	HGD (0.894)	Gabor, GLCM (0.815)

A few conclusions can be drawn from the observation of the results:

1. In most scenarios, combining descriptors with other clinical information significantly outperforms standalone clinical data and all standalone descriptors, with the highest performing combination not being necessarily based on the highest performing standalone descriptor;

2. The suitability of a descriptor depends on the type of lesion: when all lesions are mixed and when only masses are present, texture and shape descriptors (e.g. GLRL, HGD and HOG) lead to the highest performances, whereas regarding calcifications, texture descriptors (e.g. wavelets, GLCM, Gabor filter banks and curvelets) clearly outperform others;

3. The performance of the descriptors is dependent on the image database, for instance in the masses subset GLRL is in the highest performing group of the DDSM database and not in that of the BCDR database, and the contrary is observed with HOG. However, some descriptors maintain their good performances transversely to the database, as HGD in this situation.

Tahmasbi et al. [14] used Zernike moments for the extraction of shape and margin properties of lesions, with the aim of discriminating between benign and malignant lesions. Patches were manually segmented from mammograms extracted from the Mammographic Image Analysis Society (MIAS) [15] database. These were then processed differently for the computation of shape and margin features: for the computation of shape features, lesions were binarized; for the enhancement of margin properties, histogram equalization was performed, increasing the contrast of the patch. Then, low- and high-order Zernike moments were computed from each processed version of the patch and the magnitudes (invariant to rotation, as opposed to phases) were used as the feature vector. For classification, a multilayer perceptron was selected and TPR, TNR, AUC and accuracy, were used as performance measures.

It was observed that the best performance belonged to the systems that used only shape features, particularly those based on low-order Zernike moments (achieving FPR of 11.13 %, FNR of 0.0%, Accuracy of 92.8% and AUC of 0.975) , and that, as the proportion of margin features increased, the performance decreased. Hence,

1. The tuning of the parameters of a descriptor (if they exist), in this case the order of the Zernike moments, can be determinant in the performance, as features based on low-order Zernike moments outperformed those based on high-order ones;

2. The processing of the images to enhance certain properties can also be determinant, as in this case the shape-enhancing processing led to higher descriptor performances than the margin-enhancing processing.

Sharma and Khanna [16] also used Zernike moment magnitudes for the detection of malignant lesions. The patches were manually segmented from mammograms of Image Retrieval in Medical Applications (IRMA) and DDSM databases. Besides Zernike moments, two standard image descriptors, GLCM and Discrete Cosine Transform (DCT), were computed from the patches. Then, statistics were extracted from the outputs of the descriptors to build the feature vectors. For classification, kNN and SVM were selected, and sensitivity and specificity were adopted as performance measures. With SVM, performance increased up to order 20, and decreased for higher order moments, on both databases (sensitivity = 0.99 and specificity = 0.99 for IRMA database and sensitivity = 0.96 and specificity = 0.96 for the DDSM database, with order 20). kNN generally performed poorer than SVM, and needed much higher moment order to achieve similar performances (at order 35, sensitivity

= 0.97 and specificity = 0.92 for IRMA database and sensitivity = 0.94 and specificity = 0.93 for the DDSM database).

Compared to GLCM and DCT-based features, the Zernike-based descriptor outperformed both (sensitivity values of 0.90, 0.78 and 0.99 and specificity values of 0.93, 0.78 and 0.99 for GLCM, DCT and Zernike-based features respectively, using SVM and IRMA database).

Thus,

1. Zernike descriptors can outperform well-known descriptors as GLCM and DCT in malignancy classification of mammographic patches, achieving very high performances;

2. An optimizing classifier should be selected – in this case SVM was clearly advantageous compared to kNN, achieving higher performances with lower dimensionality feature vectors.

3.2 PET images

The basics of PET have already been explained in section 1.1. Many studies have been developed for the automated analysis of pathologies typically covered by PET. FDG-PET, in particular, is the most studied, as it allows for the analysis of prominent pathologies characterized by regions of abnormally high (e.g. tumors or infections/imflammations) or low glucose metabolism (e.g. Alzheimer's disease). For this section, we selected three representative studies, whose main features are summarized in Table 3.5.

Table 3.5: Main features of the studies reviewed in section 3.2.

Authors	Objective	Dataset	Descriptors Families
Wu, Khong and Chan [20]	Automatic detection and classification of nasopharyngeal carcinoma, combining image descriptors with a priori clinical knowledge.	25 PET/CT examinations of patients suffering from NPC from the PET/CT Unit in the University of Hong Kong.	Intensity Texture (second-order moments) Shape (area, eccentricity, compactness)
Naqa et al. [21]	Predictive model of treatment outcome based on intensity, texture and shape features of pre-treatment PET images.	14 images of patients with cervix tumor (7 with persistent disease after treatment) and 9 images of patients with head and neck tumors (4 of them died after treatment)	Intensity (SUV-based) GLCM Shape (compactness, eccentricity, extent, Euler's number).
Morgado [22]	Comparison of feature extraction and feature selection techniques for automated diagnosis of MCI and AD using PET images.	59 PET images of each class (healthy patients, MCI patients, AD patients) retrieved from the ADNI database [23].	Intensity Local Variance LBP

Wu et al. [20] designed a system for the detection of primary tumor and metastasis of nasopharyngeal carcinoma in PET/CT images. The patches were segmented automatically

using a region-growing algorithm. Then, texture (second-order texture moments) and shape (compactness, area, eccentricity) features were extracted.

As a way to discard false positives, i.e., physiological high-uptake spots, extra features were computed: the average intensity of CT values was used to differentiate tumor regions from regions of physiological marrow uptake in bones and brown fat uptake in fatty tissues; the intensity difference between the regional peak and its surroundings was used to differentiate regions of true bone metastasis from regions of normal bone marrow uptake; anatomical location information was also used, as the likelihood of a segmented candidate to be a part of the primary tumor or its nodal metastasis differs according to its anatomic location (by definition the primary tumor arises from the nasopharynx and a pattern of nodes spreads in the neck); the symmetry of the segmented candidate about the medial plane was also considered (symmetric organs as tonsils, salivary glands and thyroid, naturally show high uptake). Different combinations of all features were input to a SVM.

With regard to the performance of the region-growing algorithm, all lesion segmentations overlapped at least 80% of the corresponding volumes identified by radiologists. However, five false positives were segmented as well.

Regarding classification, samples from 20 PET/CT volumes were used for training; the image feature combination of relative position, average intensity, area, eccentricity and symmetry had the higher TPR (99.3 %) and the lowest FPR (4.8%), with the relative position being the most important feature, since eliminating it from the feature vector clearly worsened performance more than eliminating any other feature. Applying this model to the samples from the remaining 5 PET/CT volumes, a TPR of 95.1 %, a FPR of 7.0 % and an accuracyof 93.3 % were obtained.

In conclusion, the work by Wu et al. shows that combining image descriptors as area and eccentricity with other clinical information as average CT values, symmetry measures to the medial plane, and anatomical location can significantly improve the performance of tumor detection on PET, helping to differentiate physiological from pathological uptakes.

Naqa et al. [21] used logistic regression to build a predictive model of treatment outcomes of cervix and head and neck cancers, based on image features of the pre-treatment PET images. Lesion volumes were segmented automatically using a region-growing algorithm in the case of cervix tumors, and manually in head and neck cancer situations. SUV-volume histograms, GLCM and shape features such as Euler's number, eccentricity, extent and solidity were computed from the segmented volumes. Features computed from descriptor outputs were ranked according to Spearmann coefficient and AUC, regarding the feature-label vector relation. The top two highest scoring features for each pathology were selected as variables of the logistic regression model. Regarding the cervix tumor patients, the difference between the fractional volumes above 90% and 10 % of the maximum SUV and energy achieved the highest scores; regarding patients with head and neck tumors, the fractional volume above 90% of maximum SUV and the shape extent were the most discriminative features. For both tumor types, logistic regression analysis was used to obtain predictive models of the outcomes based on the top two discriminative features. When compared with the real outcomes, the cervix tumor model obtained a Spearman coefficient of 0.49 and an AUC value of 0.76, whereas the head and neck tumor treatment outcome predictive model obtained a Spearman coefficient of 0.89 and an AUC value of 1.0. Hence, attractive results were obtained with simple two-feature logistic regression-based predictive models of treatment outcomes, particularly in the head and neck cancer. Texture, shape and SUV-volume features turned out to be more important in discrimination than the usual SUV

statistics.

Morgado [22] compared feature extraction and selection techniques for the automated diagnosis of Alzheimer's Disease (AD), using brain PET. PET volumes of patients with AD, Mild Cognitive Impairment (MCI) – a syndrome that is proved to be related with the pre-clinical stage of AD – and also from Cognitive Normal (CN) patients, were extracted from the ADNI [23] database. Four descriptors were computed, from the entire brain volumes: voxel intensities, local variances, 2D-LBP and 3D-LBP. As there was no segmentation, descriptor outputs have very high dimensionalities. For that reason, feature selection algorithms such as Point Biserial Correlation Coefficient (PBCC), Mutual Information Maximization (MIM) and Minimal Redundancy Maximal Relevance (mRMR) were applied. A SVM was selected for classification, and accuracy was used as performance measure. The best feature – selection algorithm combination was:

1. 3D-LBP and PBCC for the AD vs. CN task (91.4 %);
2. Voxel intensities and MIM for the MCI vs. CN task (79.4 %);
3. Local variances and MIM for the MCI vs. AD task (73.4 %).

However, other combinations reached similar performances. From the results, one may observe that the most accurate classification task was AD vs. CN, with a clear drop in performance in the tasks that involve MCI. Over all the classification tasks, MIM was most frequently the best selection algorithm, followed by PBCC with similar results and mRMR with generally poorer results. Regarding feature types, it is not so easy to choose one that consistently had the best performance over all tasks. In conclusion, all voxel intensities, local variances, 2D-LBP and 3D-LBP, performed well in the diagnosis of MCI and AD, with the PBCC and MIM feature selection algorithms allowing high performances after dimensionality reduction.

3.3 CT images

CT is mostly used for the analysis of bone and lung pathologies. Bone regions stand out as very intense regions when compared to the remaining tissues due to the higher density and thus higher x-ray absorption. On the contrary, lungs stand out as especially dark regions due to their particularly low density (high air percentage). CT is also used for analysis of pathologies of other tissues, but is usually preferred over by MRI due to the high soft tissue contrast the latter provides.

Most studies using CT images for automated diagnosis are related to lung diseases, particularly lung nodules, although other pathologies such as liver pathology or polyps in CT colonography are often addressed. Lung nodules are clearly seen in CT images, as they are dense regions and contrast with the dark lung background. In fact, the high CT resolution allows for the identification of very small nodules, which sometimes is not possible using MRI. However, when one's Computer-aided Diagnosis (CAD) system is sensitive to very small nodules, it gets more susceptible to false positives, as noise or small artifacts can be considered nodules. For this reason, there are several studies on false positive reduction strategies. For this section, we selected three representative studies, whose main features are summarized in Table 3.6.

Table 3.6: Main features of the studies reviewed in section 3.3.

Authors	Objectives	Datasets	Descriptors Families
Depeursinge et al.[24]	Texture descriptors for classification of 4 different pathologies of lung tissue.	77 samples of healthy tissue, 72 of emphysema, 155 of miconodules, 64 of fibrosis and 113 of ground glass nodules from an internal database.	Gray-level histogram Wavelet
Boroczky and Zhao [25]	Finding the optimal feature subset, out of a pool of texture, shape and intensity features, for false positive reduction of a previous lung nodule detection system.	52 true positives and 443 false positives output by a previous CAD system.	Gray-level matrix Gray-level histogram Gradient matrix Shape (sphericity, enlongated and flat shape, compactness).
Dettori & Semler [26]	Comparison of texture descriptors for classification of tissues of different organs in CT scans.	2 healthy chest and abdomen CT studies from Northwestern Memorial Hospital.	Wavelet Ridgelet Curvelet GLCM GLRL

Depeursinge et al. [24] explored texture descriptors for the discrimination among 5 classes of lung tissue – healthy, emphysema, ground glass nodules, fibrosis and micronodules. Patches were manually segmented from CT images and the gray-level histogram and wavelet descriptors were computed. Three feature vectors were considered: histogram statistics, wavelet statistics and a combination of the two. A kNN classifier was selected and the leave-one out approach was used for validation. Accuracy was used as performance measure.

The following experiments were performed: 5 classification tasks of each tissue type vs. the remainder, and a multiclass task. The combination feature vector systematically outperformed the isolated ones, with accuracies in the range 95 % – 100 % with regard to the each vs. the remainder tasks and, regarding the multiclass task, 92.2 % of the healthy samples were correctly classified, all emphysemas, 86.7 % of the ground glass lesions, 92.9 % of the micronodules and 93.8 % of fibrosis.

In summary, a combination of gray-level histogram statistics and wavelet features can achieve high performances in the detection and discrimination between lung pathologies.

Boroczky and Zhao [25] search for the optimal feature subset, out of a pool of texture, shape and intensity features, for false positive reduction of a previous lung nodule detection system. Lesion volumes were manually segmented. From each segmentation, the following features were extracted: statistics computed directly from gray-level vector of intensities and from the gray-level histogram, the difference between the mean values of the gray-

levels within the nodule and in its vicinity, statistics computed from the gradient matrix, and shape features such as spheric, flat, and enlongated shape, sphericity and compactness. Then, GA were used for finding the optimal subset size and the optimal subset of features: a hierarchical fitness function was used, with the first priority assigned to the sensitivity, the second to the specificity and the third to the number of features in the subset. A first GA run was performed to determine the optimal subset size, based on the occurrences of chromosomes representing each subset size – those with 10 features were clearly the most frequent. Subset size was then fixed to 10, and a second GA run was performed to determine the optimal 10-feature group.

The fittest chromosome corresponded to the following subset: gray level minimum, compactness, flat shape, elongated shape, sphericity, contrast, gradient maximum, gradient standard deviation, gradient skewness, and gradient small value ratio. Using this feature subset, all the true positives were retained (sensitivity of 100%), and a 50% reduction of false positives was achieved.

Detori and Semler [26] compare the performance of wavelet, ridgelet, and curvelet texture descriptors, as well as two standard texture descriptors – GLCM and GLRL – in the discrimination between tissues of five different organs – spleen, backbone, heart, liver and kidney). Patches were segmented from CT images using an Active Contour algorithm. Then, descriptors were computed and statistics were extracted from their outputs. Different combinations of such statistics were built for each descriptor and input to a decision tree.

In Table 3.7 are the performances of the best feature vector of each descriptor. Observing the results, one can see that the best curvelet-based feature vector clearly outperforms the best wavelet and ridgelet-based feature vectors. Regarding the comparison with standard texture descriptors GLCM and GLRL, one can conclude that the the best wavelet and ridgelet feature vectors are outperformed by GLCM and GLRL-based ones. On the other hand, the best curvelet feature vector outperforms both. In conclusion, curvelet-based image features can be very powerful in the discrimination of textures of different tissues on CT images, outperforming in many scenarios other standard texture descriptors as GLCM and GLRL and wavelet and ridgelet-based features.

Table 3.7: Average performances over all organs of the highest performing feature vectors within the wavelet, ridgelet and curvelet groups and of the GLCM and GLRL.

Descriptor	TPR (%)	TNR (%)	Prec. (%)	Acc. (%)
Wavelet (Haar)	74.4	93.7	74.4	89.9
Ridgelet	83.8	96.0	85.0	93.6
Curvelet	94.6	98.7	94.7	97.9
GLCM	89.1	97.3	89.7	95.8
GLRL	84.3	96.1	84.7	93.9

3.4 MR images

Magnetic Resonance (MR) images are based on the different relaxation times of tissues after being subjected to an electromagnetic stimulus, and provide excellent soft tissue contrast. The relaxation time of a tissue is related with its water content. Moreover, contrast can be further enhanced through the injection of a contrast enhancement agent. For this

reason, MRI is widely applied in the diagnosis and treatment of neurological, cardiovascular, musculoskeletal, liver and gastrointestinal diseases.

Most automated diagnosis studies using MR images are related to brain pathologies as tumors, dementia, or lesions related to White Matter Hyperintensities (WMH) – these have shown to be associated with several prominent pathologies, such as multiple sclerosis, vascular disease and dementia, although other pathologies as breast or prostate cancer are also frequently object of study [27] [28]. We selected three representative studies, whose main features are presented in Table 3.8.

Table 3.8: Main features of the studies reviewed in section 3.4.

Authors	Objectives	Datasets	Descriptors Families
Unay, Ekin, Cetin, Jasinschi & Ercil [29]	To demonstrate robustness of LBP texture descriptors to bias field and rotation degradations.	Dual (T2 and PD) MR scans from 549 subjects from Leiden University Medical Center.	LBP
Reddy, Solmaz, Yan, Avgeropoulos, Rippe & Shah [30]	To prove that incorporating the information of a confidence surface in segmentation methods, built over the output scores of a brain tumor classifier, significantly improves their performance.	19 groups of MRI images with brain tumor.	Intensity HOG LBP
Theocharakis et al. [31]	Discrimination between two white matter hyperintensities-related lesions, multiple sclerosis (MS) and cerebral microangiopathy (CM)	47 CM and 31 MS ROIs of MR images of an internal database.	Gray-level histogram GLCM GLRL

Until recently, analysis of brain MR images was mostly exclusively based on intensity features. Since soft tissue contrast is high, intensity features are naturally discriminative. However, if our aim is brain lesion segmentation for accurate tumor volume measurement, intensity-based analysis may not be enough: bias-fields (intensity inhomogheneity caused by equipment interferences during acquisition) and inter- and intra- patient misalignment significantly degrade the performance of automatic segmentation techniques. Moreover, normal tissues may also be enhanced with contrast agent, resulting in the segmentation of a larger region than the actual lesion; on the other hand, the presence of noise or non-uniformity of the distribution of contrast agent in the lesion may result in an incomplete extraction.

Unay et al. [29] suggested a LBP-based texture analysis as a potentially more robust complement or alternative to intensity-based analysis. In order to test robustness of LBP to bias fields and rotation degradations, original MR images were degraded using a set of simulated bias fields (with larger or smaller intensity and spatial variations) and rotated by several angles using three different interpolation methods. Then, some LBP variants were

computed from the original and degraded images. The Bhattacharyya distance between each descriptor computed in the original and the degraded image was used as the dissimilarity measure.

It was observed that dissimilarity increased with the increase of intensity and spatial variations of bias fields. In the case of rotation degradation, dissimilarity was higher for large rotation angles and for lower-complexity interpolation methods. Despite these increases, all dissimilarity values fell below 0.04% in the case of bias field degradation and bellow 4% in the case of rotation degradation. Regarding the different variants of LBP, introducing rotation invariance and uniformity in LBP increased performance.

In conclusion, the uniform and rotation invariant LBP variant is quite robust to bias-fields and rotation degradation, and thus a promising complement to the usual intensity-based analysis.

Reddy et al. [30] also believe that it is advantageous to add texture features to the usually used intensity features on MRI, based on the fact that normal brain tissues differ also in structure from lesions. They propose confidence-guided versions of known segmentation methods and compare their performances with those of the original versions, in a brain tumor segmentation problem.

To start, a mask for the enhanced region is generated with the difference between T1pre and T1post images.

Then, mean intensity, LBP and HOG features are computed for each pixel within the enhanced region mask from each of T1pre, T1post, T2 and FLAIR images, and concatenated to form a single feature vector.

After that, the feature vectors of each pixel are input into two different classifiers, SVM and AdaBoost, for tumor pixel classification. A confidence surface is then constructed based on the classification output scores. The authors propose to use the generated confidence surface to guide the segmentation process: two classical segmentation methods, level set and region growing, are slightly modified to incorporate the confidence surface information in the segmentation process.

Regarding classification results, Receiver Operating Characteristic (ROC) curves were plotted and it was observed that AdaBoost outperformed SVM, with larger AUC values. Checking the AdaBoost weights for the different features, the MI features from T1pre and T1post had the larger weights, indicating that these features still play the most important role in tumor detection. On the other hand, the HOG features from T1pre and T1post images and the LBP feature from T1pre had larger weights than the mean intensity feature from T2 and FLAIR images, suggesting that these texture features are also useful for discrimination.

For assessing segmentation accuracy, the average Dice Similarity Score (DSS) was computed. Using the original level set method, a DSS of 0.3 ± 0.27 was obtained, whereas for original region growing DSS value was 0.29 ± 0.22. Using the confidence guided versions, DSS significantly improved for both methods, with confidence guided region growing segmentation outperforming confidence guided level set segmentation: DSS = 0.68 ± 0.13 for level set and 0.69 ± 0.14 for region growing.

Thus, it can be concluded that intensity features are still probably the most important for brain tumor analysis, but SVM and HOG features also contribute for discrimination. Moreover, incorporating confidence guiding in the segmentation methods significantly improved their performances.

Theocharakis et. al [31] developed a system for discriminating between Multiple Sclerosis (MS) and Cerebral Microangiopathies (CM), based on texture features.

Patches were manually segmented and gray-level histogram, GLCM and GLRL descriptors were computed. Statistics computed from descriptor outputs were extracted. For feature selection and classification, four methods were compared: minimum distance classifier, linear discriminant analysis, logistic regression and PNN. With a leave-one-out cross-validation approach, the best classification accuracy occurred with PNN (88.46 %), using the mean value, sum of variance and run-length nonuniformity features. However, a cross-validation scheme with $\frac{2}{3}$ train – $\frac{1}{3}$ test dataset partition led to an average accuracy of 72.96 % (over 10 random partitions), with different features in the top 3 for each of the repetitions. The most frequent features on the top 3 were mean, contrast, sum of average and sum of variance. A total of 15 of the 23 features were at least once in the top 3. A Mann-Whitney U-test was performed to assess significant difference between both classes for each of these 15 features, and 13 of them showed significant difference.

Thus, a combination of features computed from the gray-level histogram and the GLCM and GLRL matrices can lead to relatively high performances in the discrimination between MS and CM.

3.5 Artificial Neural Networks – Revised Works

ANNs have a rich history concerning medical image analysis applications, namely in segmentation and classification tasks.

Regarding segmentation, unsupervised or clustering neural networks are normally preferred. As our goal is classification, we will not further explore such subject.

Verma and Zakos [32] used a MLP with one 10-unit hidden layer for discrimination between benign and malignant breast microcalcifications. With only three inputs - two gray-level features and the number of pixels -, computed from microcalcification regions in mammograms, an accuracy of 88.9% was achieved.

Halkiotis et al. [33] aimed for the detection of clustered breast microcalcifications. Five features were computed from the candidate Region of Interest (ROI)s, four of them being moments computed from the gray-level histogram and the remaining being the number of objects in a limited neighborhood. These features were input to a MLP with one 10-unit hidden layer for classification as either clustered microcalcification or not. A sensitivity of 94.7% was obtained, with 0.27 false findings per image.

Papadopoulos et al. [34] developed a system also for the detection of microcalcification clusters. From mammogram ROIs, 22 intensity, shape and texture cluster features were extracted. Features of candidates that passed a rule system aiming at false positive elimination were subjected to principal component analysis for posterior dimensionality reduction. Dimensionality was reduced to 9, through the elimination of components that contributed with less than 3% of the overall data variance. Then, the 9 features were fed to a two-hidden-layer MLP, the first layer having 20 units and the second having 10. An AUC value of 0.91 and 0.92 were obtained for two different datasets.

Christoyianni et al. [35] used an RBFNN classifier, first for detecting abnormal breast tissue and, secondly, to discriminate between benign and malignant breast lesions. Three feature vectors were compared: one consisted of gray-level histogram moments, a second one was composed of statistics computed from the co-occurrence matrix and the third was composed of the principal components (5 in the case of abnormality detection and 8 in the case of dicrimination between benign and malignant) of the coefficients resulting from

Independent Component Analysis (ICA). The ICA-based feature vector was the highest performing one in both classification tasks, with an accuracy of 88.23% in the abnormality detection scenario and of 79.31% in the benign from malignant discrimination.

Chen et al.[36] used a PNN for discrimination between two types of liver tumors hemageoma and hepatoma, based on three image features computed from CT ROIs - correlation, sum entropy and normalized fractional brownian shape. An 83 % accuracy was obtained.

Some network models are designed to operate directly on images. Examples are convolutional neural networks - used, for instance by Sahiner et al. [37] for detection of abnormal breast tissue in mammogram ROIs, achieving an AUC of 0.87 – and massive training artificial neural networks – applied by Susuki et al. [38] in lung nodule detection and false positive reduction, maintaining a sensitivity of 96.4% and a achieving false positive rate reduction of 33% when compared to a previous system designed by the same authors.

Pruning algorithms can be used for the elimination of network connections considered irrelevant - in the sense that their presence does not affect performance - during training (e.g. input selection). Setiono [39] applied a pruning algorithm to a MLP in a breast cancer diagnosis application.

ANN ensembles have also been explored, with the final output being a function of the outputs of different ANNs. Zhou et al. [40] used an ANN ensemble for lung cancer detection.

3.6 Genetic Algorithms – Revised Works

GAs have already been used in a number of applications related with medical image analysis, namely in the optimization of segmentation, feature selection and classification methods.

The problem of medical image segmentation can often be expressed as one of optimization of an objective function. For example, Active Contour Model or Snake segmentation [41] is a popular segmentation method based on energy minimization of the so-called snakes. Snakes are continuous splines under forces of three natures:

1. *External constraint forces* These forces place the snake near the wanted boundary. Usually, this is done by manually tracing a spline near the boundary.

2. *Internal forces* These forces impose a piecewise smoothness constraint.

3. *Image forces* These forces push the snake toward salient image features such as lines or edges.

Thus, a draft of a snake is first drawn near the boundary, then the internal and image forces will reshape the snake so as to smoothly adapt to the boundary. This happens as a consequence of a condition of energy minimization of the snake – Active Contour segmentation is a problem of minimization of an energy objective function. Mathematically, it can be expressed as:

$$min\left(E_{snake}\right) \leftrightarrow min\left(E_{external} + E_{internal} + E_{image}\right)$$

As powerful function optimization tools, GAs can be used for such task. For example, they have been used in the segmentation of anatomical structures in ultrasound [42] and in the segmentation of the foveal avascular zone in retinal images [43].

Another popular segmentation method is the template-matching method. In this method, the detection of an object is usually based on a correlation metric to a template object. GAs have been introduced in such methods as a way to detect objects by maximizing a correlation-based fitness function. For example, Lee et al. [44] use GAs in a template-matching technique for the detection of pulmonary nodules in Helical CT images. As template nodules, they used simple models that simulate real nodules – they observed that CT values in nodules followed a gaussian distribution, and built 4 gaussian models with 4 different diameter values – within the 5-30 mm interval (most nodules fall in this size range). Chromossomes of 25-bit length were built – 23 bits containing the (X,Y,Z) coordinates of the candidates, and 2 bits for the selection of the reference image (i.e. 1 of the 4 gaussian models). The fitness of a chromossome is the cross-correlation coefficient between the region centered at the coordinates and the selected reference image. Chromossomes with fitnesses over a certain value were considered nodule candidates. Then, some features were computed in order to eliminate false positives: mean and standard deviation were used to discard false positives in bone, skin and mediastinium, since these usually show higher CT values than lung nodules; area, circularity and irregularity measures were also used to discard some false positives; contrast, maxmean CT value and gradient direction were used to discard blood vessels running vertically regarding the slice image. In the experiment, 20 clinical cases were used – 15 abnormal and 5 normal cases. A total of 98 nodules were detected by radiologists in the abnormal cases. The genetic algorithm template matching technique was able to detect 55 of the 98 nodules with 3224 false positives; after the feature-based false positive elimination stage, the number of false positives went down to 333 (a 88% reduction) and also a true positive was mistaken for an artifact, lowering the true positives to 54. Thus, a 72% sensitivity was achieved.

Ye et al. [45] also use a Generic Algorithm Template Matching (GATM) algorithm for lung nodule detection in CT images, but with phantom nodule images instead of gaussian models.

Another common application of GAs is feature selection. Usually, chromossomes have lengths equal to the total number of features – one bit per feature – and the bit is 0 or 1-valued depending on whether the corresponding feature is selected or not. The fitness function of an individual is usually based on one or more performance measures of the classifier (e.g. AUC, sensitivity, specificity) and sometimes a penalty term proportional to the number of 'active' features. Herein we present some examples:

Sun et al. [46] apply genetic algorithms for feature selection in the detection of breast cancers in mammograms. The dataset consisted of 164 cancer regions and 296 normal regions extracted from mammograms of the DDSM database. Descriptors such as GLCM, Gabor filters, wavelets and curvilinear structure were computed from the patches , and 86 features were extracted from their outputs. These 86 features were used to build 86-bit chromossomes – 1 bit per feature – with 1 or 0 values depending on whether the corresponding feature is selected or not. The fitness value of each chromossome was the area under the ROC curve produced by a classifier based on linear discriminant analysis. Two genetic algorithm models were used – a traditional genetic algorithm, and the CHC model – achieving AUC values of 0.903 and 0.932, respectively.

Sahiner et al. [47] use genetic algorithms for feature selection in the detection of masses in mammograms. The dataset included 168 mammograms, 85 of which containing benign masses and 83 containing malignant masses. From each mammogram, 4 ROIs were extracted, one containing the mass, and 3 others containing normal breast tissue. Of the 3 normal tissue

ROIs, one should contain dense tissue which could mimic a mass lesion, another would be a mixture of dense and fat tissue and another mainly showing fat tissue. From each ROI, GLCM and wavelet-based features were computed. For ROI segmentation, they used a pixel-by-pixel clustering algorithm followed by binary object detection. After detecting a single suspicious object within each ROI, shape and margin features were extracted from it – e.g. perimeter, area, circularity, rectangularity, contrast, perimeter-to-area ratio, and radial length features. Chromossomes were built with a bit per feature – 1 if the feature was selected and 0 otherwise. The fitness function was made of a main component – the area under the ROC curve of a Fisher's linear discriminant classifier – and a penalty term – proportional the number of features. An average AUC of 0.90 for 20 features was obtained with GA-based feature selection, outperforming random and stepwise Linear Discriminant Analysis (LDA) selection techniques.

Sometimes GAs are used to find the operating parameters of classifiers that optimize their performances. Usually, the parameters to tune are encoded in the chromossomes and the fitness function is based on one or more performance measures of the classifier (e.g. sensitivity, specificity). Herein we show examples regarding simple threshold-based classifiers and more complex ANN classifiers.

Anastasio et al. [48] use genetic algorithms for the optimization of the detection of clustered microcalcifications in mammograms. 89 mammograms were used, 82 of which containing clusters of microcalcifications. First, the original mammogram was pre-processed by linear filtering in order to increase the signal-to-noise ratio of the microcalcifications. Then, the potential microcalcifications were identified by gray-level thresholding and morphological filtering. After that, a set of features was extracted in order to eliminate false positives. The elimination was based on 10 thresholds of intensity, power spectrum, contrast, linearity and area. The genetic algorithm was used for finding the optimizing values for those thresholds – a chromossome was made of 10 values. The cost function was based on a sensitivity-specificity tradeoff – higher costs were assigned to less desirable sensitivity-specificity pairs and vice-versa. Bilinear interpolation was used for obtaining the cost associated with the non-tabulated values of sensitivity/specificity. An 87% sensitivity was obtained at 1.0 false positives per image.

GAs have also been used for determining the optimizing parameters of classifiers. For instance, Neuroevolution is a field of machine learning that uses GAs to evolve ANNs – one does not need to propose the network parameters because a method called NeuroEvolution of Augmenting Topologies (NEAT) [49] automatically discovers the best network topology and weights that best fit the complexity of the task. Tan et al. proposed a feature (de)selective version of NEAT – FD-NEAT [50] –, where an additional mutation operator enables discarding irrelevant or redundant inputs. Tan et al. have been applying FD-NEAT in works related to CT lung nodule detection [51],[52].

Chapter 4

Experimental Setup

In this chapter, the adopted methodology and parameterization for the data processing, classification and regression stages is reported.

4.1 Data

We were provided clinical data regarding 63 patients suffering from two oncological diseases of distinct natures, with 29 of them suffering from Hodgkin Lymphoma and the remaining 34 suffering from neuroendocrine tumors. Clinical data of each patient was collected by a nuclear medicine team of IPO-Porto, and includes:

1. PET/CT exams of the patient before and after radiotherapy treatment;
2. Patient age and tumor stage by the time of the first PET/CT exam;
3. Patient weight before and after treatment;
4. Maximum SUV within the main lesion and maximum SUV of a reference organ, before and after treatment (the reference organ is useful for monitoring any procedural mistake that would lead to the wrong SUV estimation – if the SUV of this organ shows unrealistic values, then the lesion's SUV must not be trusted either).

The development of multiple lesions is quite common in patients suffering from both pathologies. However, we only have access to maximum SUV information of the main lesion, i.e., the lesion of highest uptake. Thus, we will only follow one lesion per patient, the main lesion. Four types of lesion response-to-treatment are observed in our dataset, each with the following number of instances:

1. Negative – the lesion has become more malignant – 2 instances. The leave-one-out cross-validation approach was selected with the aim of assuring that, in the experiments with the original dataset, there would always be at least one sample of the negative response class in the training set;
2. Neutral – the lesion did not respond to treatment – 6 instances;
3. Positive (partial) – the lesion has decreased the malignancy degree after treatment – 27 instances;

4. Positive (complete) – the lesion disappeared after treatment – 28 instances. Because some lesions disappear, it does not make sense to compute shape features from the patches, such as area, perimeter, and so on.

4.2 Segmentation

The PET/CT information provided to us consisted of the axial slices of whole-body scans. We found the visual detection of lesions in coronal context more intuitive; hence, for segmentation, we grouped the slices into volumes and extracted the coronal slices. The segmentation process of a lesion required three main steps:

1. Identifying the main lesion – many patients present multiple lesions. Our interest is to keep track of the main lesion, i.e., the one that shows the highest malignancy degree. In order to identify the lesion of higher malignancy we summed all slices into a maximum intensity projection image, and, in the latter, we could visualize which lesion had the highest overall uptake;

2. After identifying the anatomical location of the lesion to segment, we had to detect the slice where it showed highest uptake;

3. After identifying such slice, we proceeded to the segmentation itself. Three different segmentation processes were adopted according to the characteristics of the slice:

    a) If the intensity maximum of the slice was within the lesion, automatic segmentation was performed by applying the region-growing algorithm to the whole image.

    b) If it were on other organs, we performed semi-automatic segmentation – we drew a rectangle that enclosed the lesion, leaving out higher intensity regions, and applied region growing within that rectangle;

    c) If there were no lesion (in the cases of complete positive response), we performed manual segmentation, drawing a rectangular patch enclosing the region where the lesion used to be.

For reasons that lie with their own definitions 2.2 some of our image descriptors require rectangular patches. Thus, in cases where region growing was applied, we used the smallest rectangle to enclose the output region of the region growing algorithm, for descriptor computation.

4.3 Descriptor Computation and Feature Extraction

After a lesion was segmented, the descriptors were computed from the patch and features were extracted from their outputs, as described in Table 4.1;

Table 4.1: Parameters used for descriptor computation and features extracted from their outputs.

Descriptor	Parameters	Extracted features	Dimension
gray-level histogram	(no parameters)	mean standard deviation skweness kurtosis	4
GLCM	$\theta = 0, 45, 90, 135$ degrees $d = 1$	GLCM features	22
GLRL	$\theta = 0, 45, 90, 135$ degrees	Run Length features	11
Wavelets	2 levels of resolution Daubechies family	mean standard deviation energy entropy (for each of the 7 resulting matrices)	28
Gabor filters	3 frequencies 4 orientations (12 filters)	mean standard deviation energy entropy (for each of the 12 output images)	48
LBP	$R = 8$ block-size equal to smallest side of the patch	histogram of uniform patterns	58
			TOTAL= 171

Thus, we have 171 image features characterizing each lesion.

4.4 Final Datasets

After computing the image features of the main lesions of each patient from before and after treatment, we built a 350-dimensional feature vector for each patient, containing:

1. Before and after image features – which account for $171 * 2 = 342$ of the 350 features
2. 8 other (already mentioned) features provided by IPO-Porto:
   a) Patient age and tumor stage by the time of the first PET/CT exam;
   b) Patient weight before and after treatment;
   c) Maximum SUV within the main lesion and maximum SUV of a reference organ, before and after treatment.

To each of these feature vectors, we assigned the corresponding treatment-response class label. Thus, our dataset is composed of 63 instances of 350 features.

One is fast to notice two potentially problematic aspects of our dataset:

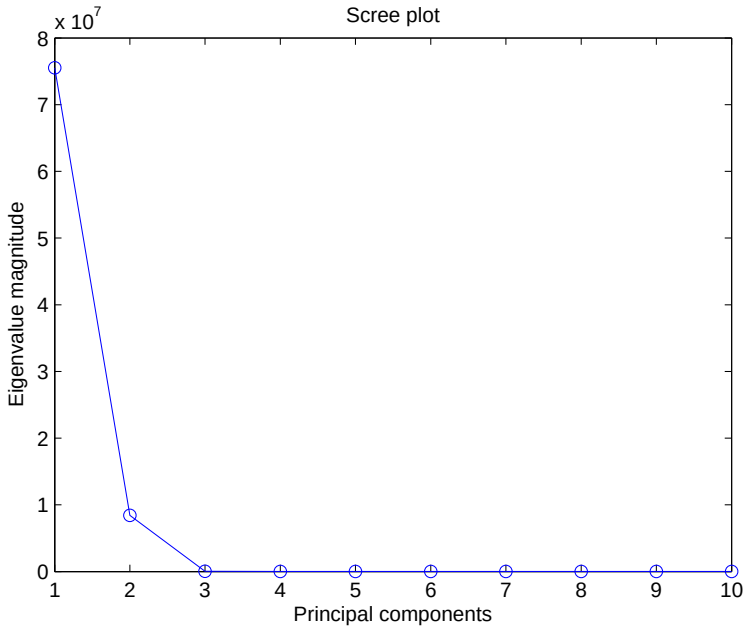


Figure 4.1: Scree plot of the first 10 principal components.

1. High dimensionality – complexity of the classification process is said to increase exponentially with dimensionality;

2. High imbalancement – not having a balanced number of instances among classes may result in poor training regarding minority classes, and a consequent tendency of bad generalization with respect to minority-class samples.

In order to assess the effect of these characteristics in the final results, we also ran our experiments after

1. Dimensionality reduction using PCA – looking at the Scree plot of our data (Figure 4.1), we would select the first two components. Moreover, as we can see on Table 4.2, the two first principal components comprehend more than 99,9% of the overall data variance. As such, we kept only two principal components;

Table 4.2: Percentage of overall variance comprised in the first three components. Percentage keeps decreasing for further components.

Principal component	Data variance percentage
1	89,92
2	10,02
3	0,040

2. Synthetic data generation using SMOTE – we generated 6 samples for each of the two minority classes (negative and neutral responses), so as to enable some more training regarding these classes.

In summary, 4 datasets will be used in the experiments, so as to analize the effect of data dimensionality and balancement in the results.

1. The original dataset;
2. The original dataset after SMOTE;
3. The original dataset after PCA;
4. The original dataset after SMOTE and PCA.

4.5 Methodology and Parameterization of the Experiments

In this section, the methodology and parameterization of the classification and regression experiments is approached.

For the classification experiments, the classifier configurations in Table 4.3 were explored, in order to find the ones which optimized performance. The selected sampling strategy was leave-one-out and the classification accuracies regarding the 4 classes were computed.

Table 4.3: Explored classifier architectures.

Classifier	Parameters	Values
MLPI	number of hidden layers	1
	number of neurons in the hidden layer:	even numbers in [6,28]
MLPII	number of hidden layers	2
	number of neurons:	
	– first hidden layer	even numbers in [6,28]
	– second hidden layer	half the number of neurons of the first layer
LVQNN	number of neurons in the hidden layer	even numbers in [6,28]
RBFNN	spread value	powers of two with integer exponents in [-1,15]
PNN	spread value	powers of two with integer exponents in [-1,15]
kNN	number of neighbors	integers in [1,30]

As for the symbolic regression experiments:

1. At first, the input dataset is partitioned in training and testing sets – 70% for training and 30% for testing.
2. Then, symbolic regression is run with the configuration parameters in Table 4.4.
3. Steps 1. and 2. are repeated 30 times, with different random seeds.

Table 4.4: Main parameters of the symbolic regression simulations.

Parameter	Value
Population size	100
Maximum number of generations	1000
Terminal set	features and random constants
Function set	+, –, /, *, *ln*, *sin*, *cos*, if-then-else, sigmoid
Operators	mutation, crossover, copy (with variable probabilities)
Fitness function	sum of absolute differences between expected and obtained outputs

We implemented an elitist algorithm, where the best individual of both parents and children is given the highest priority to enter the new population; in non-elitist algorithms, children receive highest priorities, even if they are less fit. Introducing elitism will assure that training fitness cannot decrease over the generations.

Chapter 5

Results and Discussion

In this chapter, the results of the classification and symbolic regression experiments are reported and discussed.

5.1 Classification Experiment

In order to determine the appropriate settings of each classifier, for each experiment, we ranked the tested configurations by performance for each class, and the one with the best average rank (over the 4 classes) would be the one selected. The results are exposed in Appendix A. The next step was to determine which classifiers were the best for each experiment and the respective performance. For classifier comparison, we adopted Friedman's non-parametric test, followed by the Bonferri-Dunn test.

Following the same notation of section 2.9, in our specific problem $N = 4$ (classes) and $k = 6$ (classifiers), with kNN being the baseline classifier. By looking at F distribution tables, $F(5, 15) = 2,90$. Table 5.1 show the main results for Friedman and Bonferri-Dunn tests for each experiment.

Table 5.1: Classifier accuracies (ranks) for each class, the Friedman statistic ($T1$) and average classifier rank, for the experiments with the 4 datasets.

Dataset	Classifier	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
Original $T1 = 1,956$	kNN	0 (4,5)	0,167 (5,5)	0,741 (4)	0,893 (4)	–
	LVQNN	0 (4,5)	1,0 (1)	1,0 (1)	0,929 (2)	–
	MLPI	1,0 (1)	0,667 (2)	0,778 (3)	0,893 (4)	–
	MLPII	0,5 ()	0,5 (3)	0,889 (2)	0,893 (4)	–
	RBFNN	0 (4,5)	0,167 (5,5)	0,444 (6)	0,964 (1)	–
	PNN	0 (4,5)	0,333 (4)	0,704 (5)	0,750 (6)	–
SMOTE $T1 = 12$	kNN	0,125 (6)	0,417 (6)	0,630 (4,5)	0,893 (4)	5,125
	LVQNN	1,0 (1,5)	1,0 (1)	1,0 (1)	1,0 (1)	1,125
	MLPI	0,5 (3)	0,5 (5)	0,778 (3)	0,930 (2,5)	3,375
	MLPII	1,0 (1,5)	0,917 (2)	0,852 (2)	0,930 (2,5)	2
	RBFNN	0,25 (5)	0,667 (4)	0,444 (6)	0,393 (6)	5,25
	PNN	0,375 (4)	0,889 (3)	0,630 (4,5)	0,679 (5)	4,125
PCA $T1 = 4,241$	kNN	0 (4,5)	0,167 (5,5)	0,741 (5)	0,893 (3,5)	4,625
	LVQNN	0 (4,5)	1,0 (1)	1,0 (1)	0,929 (1,5)	2
	MLPI	0,5 (1,5)	0,5 (2,5)	0,778 (3,5)	0,929 (1,5)	2,25
	MLPII	0,5 (1,5)	0,5 (2,5)	0,889 (2)	0,893 (3,5)	2,375
	RBFNN	0 (4,5)	0,333 (4)	0,667 (6)	0,464 (6)	5,125
	PNN	0 (4,5)	0,167(5,5)	0,778 (3,5)	0,821 (5)	4,625
SMOTE+PCA $T1 = 4,602$	KNN	0,5 (3,5)	0,67 (4)	0,630 (5)	0,893 (4)	4,125
	LVQNN	1,0 (1)	1,0 (1)	0,963 (1)	1,0 (1)	1,0
	MLP	0,125 (6)	0,833 (2,5)	0,704 (4)	0,964 (2,5)	3,75
	MLPII	0,5 (3,5)	0,833 (2,5)	0,815 (2)	0,964 (2,5)	2,625
	RBFNN	0,375 (5)	0,417 (6)	0,741 (3)	0,714 (5)	4,75
	PNN	0,625 (2)	0,583 (5)	0,556 (6)	0,679 (6)	4,75

We can observe that, in the experiment with the original dataset, $T1 < F(5, 15)$ so the null hypothesis cannot not be rejected, i.e., one cannot state that there is a significant statistical difference among the performances of the k classifiers. In the remaining experiments, $T1 > F(5, 15)$ – we can state that a significant statistical difference exists among the performances of the k classifiers. For these experiments, CD was computed for a 5% significance level – $CD = 1,92$. The average ranks of the classifiers were also computed. Only those with an average rank better than that of the baseline classifier by more than the CD value, can be considered to significantly outperform the baseline classifier.

Looking at Table 5.1 one can conclude that:

1. In the experiment with the original dataset after SMOTE, LVQNN and MLPII significantly outperform kNN;

2. In the experiment with the original dataset after PCA, LVQNN, MLPI and MLPII significantly outperform kNN;

3. In the experiment with the original dataset after SMOTE+PCA, LVQNN is the only classifier to significantly outperform kNN.

Thus, in 3 out of 4 scenarios it was verified that the selection of a more complex classifier than kNN, such as LVQNN, MLPI or MLPII, pays off in terms of performance. In the experiment with the original dataset, such selection seems to be unjustified.

Taking a closer look at performance itself, one can draw a few relevant conclusions:

1. Our set of features allows for very high classification performances, when data is properly balanced. That is not true for all the classifiers, but the aim of this stage was to prove that the feature set allowed for high performances, the classifiers themselves not being the scope. This was, in fact, the reason why a few classifiers were used as opposed to only one, i.e., to avoid that a bad classifier choice would dictate our conclusions.

2. In data imbalancement scenarios, performance is clearly poorer – the introduction of SMOTE markedly improves performance;

3. Dimensionality reduction to two components using PCA does not seem to have significant effects on performance. As such, dimensionality reduction is advantageous for us, as it allows for a serious reduction of computational load while preserving performance.

As for time complexity, although it is not critical in this project (as real-time is not required) an idea of its order with respect to each of the adopted classifiers can be provided: the average kNN ran in 57 seconds, RBFNN and PNN in the 1 hour order (1,4 h and 0,7 h), LVQNN, MLPI and MLPII (the highest performing classifiers) ran for 14, 14 and 20 hour respectively. Naturally, in the PCA experiments, these times were largely reduced for seconds and minutes orders, for all the algorithms.

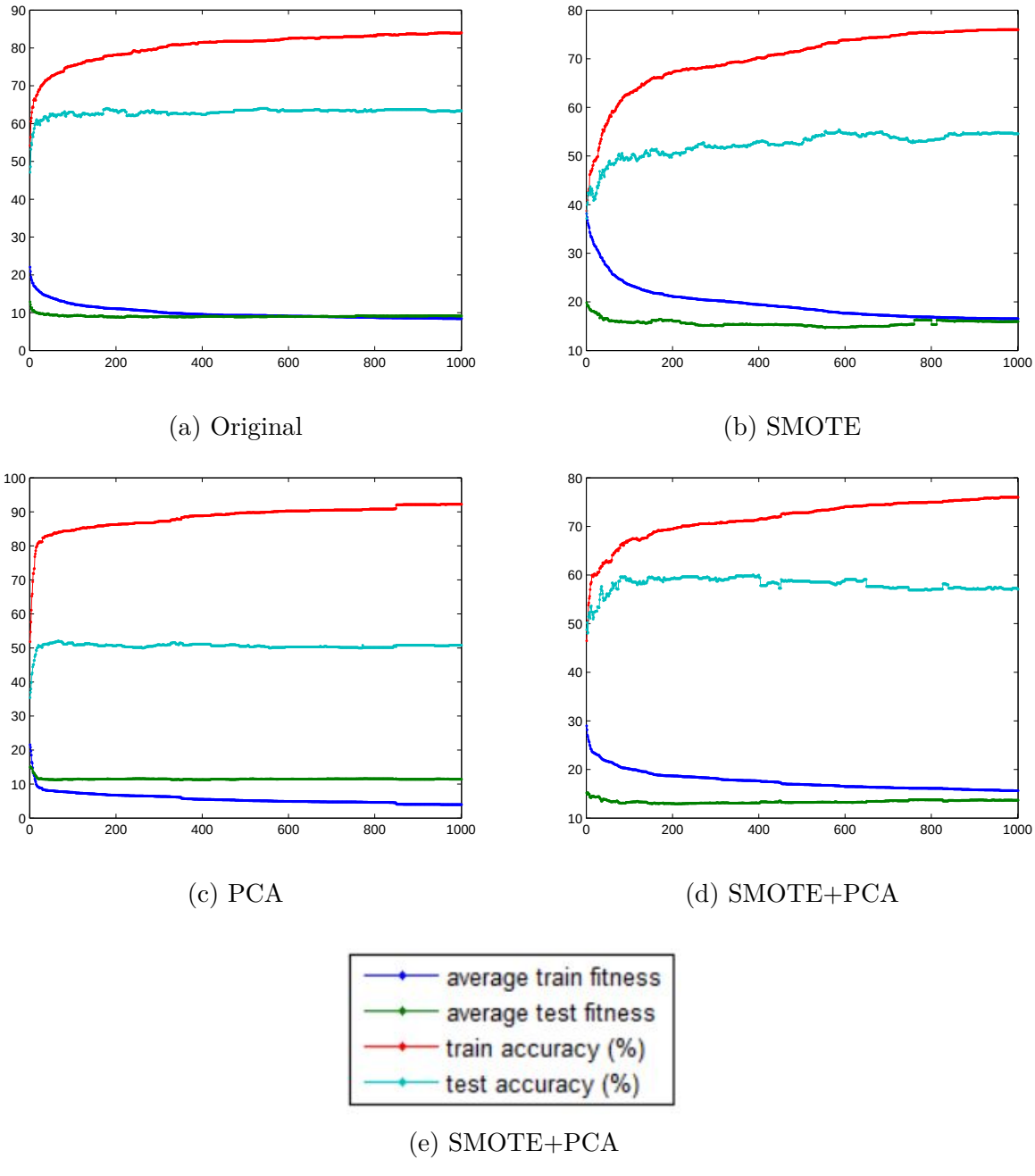
5.2 Symbolic Regression Experiment

For each of the 30 runs, of each of the four experiments, we computed the evolution of the train and test fitness, as well as of the fraction of train and test hits (accuracies), of the best individual over the generations. The average evolution (over the 30 runs) of such values is shown in Figure 5.1, for each experiment.

In the final generation, we are averaging the overall best individuals of each run (in terms of training, since testing performance does not influence the evolutionary algorithm), as a consequence of elitism. This means that at this point, every run is at its minimum training error so far. It could decrease even more if we extended the number of generations. However, minimizing training error does not always imply minimizing the testing error. If the training is exaggerated, overfitting may occur.

One may observe that neither of the four experiments stands out from the others in average performance terms, all with not very optimistic average testing performances (all fall in the 50-60% hits range). However, that may not be true in individual terms: we looked for the best performing individual out of the 30 best of each experiment. For comparing individual performances, the accuracies of the four classes were computed and averaged, so as to give equal importance to minority and majority class accuracies. This way, if the individual is very accurate for the majority classes, the performance measure will be pulled down if the same is not true for the minority classes. If overall accuracy was used, results could be deceivingly optimistic. Table 5.2 shows the accuracy information of the individuals with highest average accuracy (over the 4 classes) of each experiment. Experiments without

SMOTE are naturally harmed in the negative class, as only two samples exist. Of the four individuals, the one from the SMOTE experiment is clearly the most balanced in terms of class-wise accuracy, with reasonably high accuracies within each class – (80%, 75%, 85,71%, 88,89%). The function represented by this individual is represented in Figure 5.2. If we were to select a function at this stage, that would be the one.



(a) Original

(b) SMOTE

(c) PCA

(d) SMOTE+PCA

(e) SMOTE+PCA

Figure 5.1: Average evolution of performance for each experiment.

Table 5.2: Class-wise accuracies and their average, for the best individual of each experiment.

	Original	SMOTE	PCA	SMOTE+PCA
Average	0,60	0,82	0,42	0,62
Negative	0	0,80	0	0
Neutral	0,50	0,75	0,20	0,80
Positive (partial)	0,90	0,86	0,71	0,80
Positive (complete)	1,0	0,89	0,75	0,89

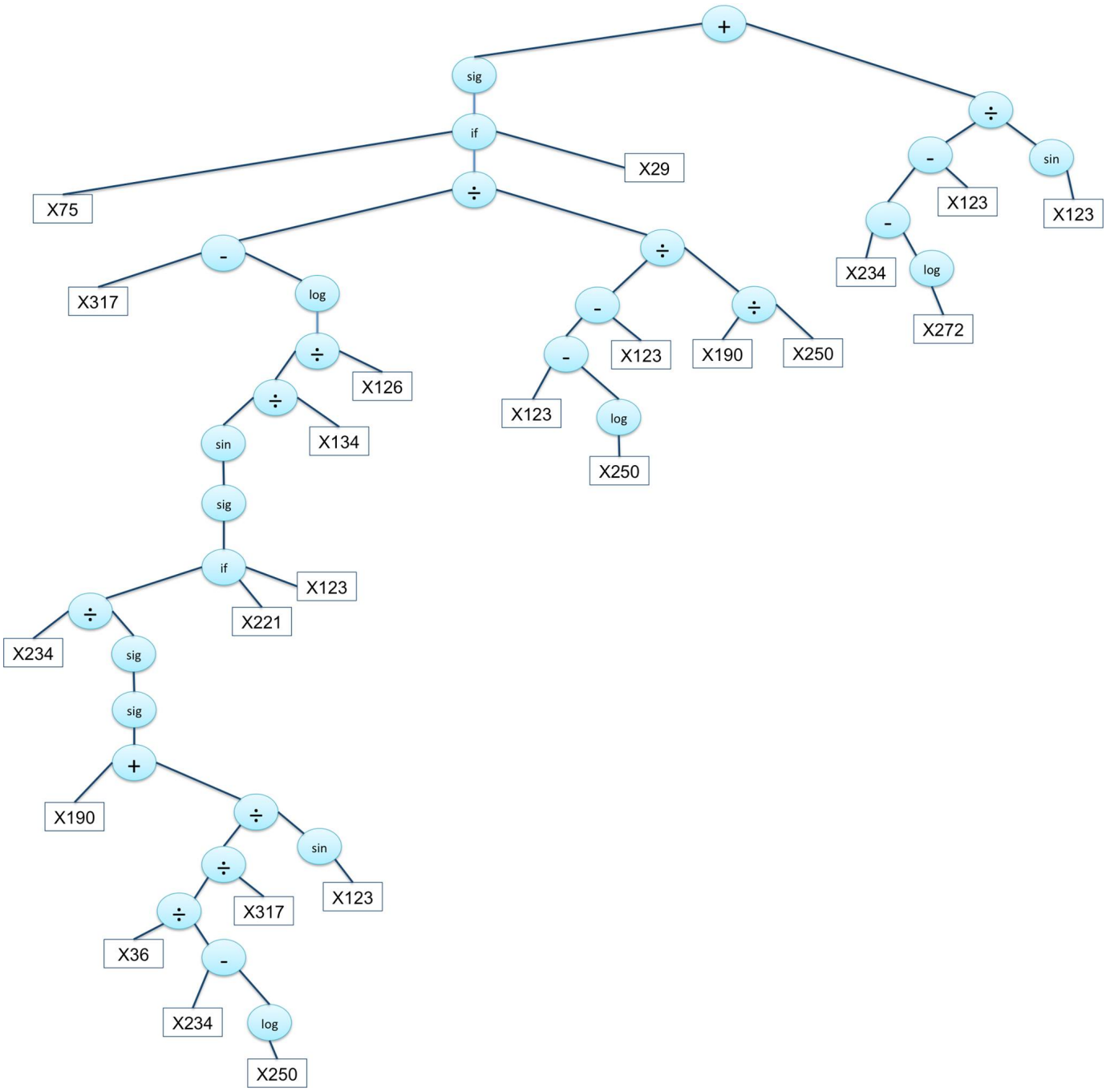


Figure 5.2: Tree representation of the best individual, i.e., best obtained evaluation function.

As a way to conclude on which are the most useful features for discrimination, an analysis of the feature selection process by the evolutionary algorithm can be performed. We estimated the percentage of individuals (out of the 30) that selected each of the 350 features of the dataset. It was observed that some features appear in several (maximum of 9) of the best individuals. The top 8 most frequently selected features are listed in Table 5.3. Just in this top, all used descriptors are represented, mixing before and after-treatment features.

Table 5.3: Top 8 most frequently selected features and the percentage of individuals which use them.

Feature Index	Related Descriptor	Before/After Treatment	Percentage
123	LBP	Before	30
250	Gabor	After	30
19	GLCM	Before	26,67
104	Gabor	After	26,67
182	Histogram	After	26,67
214	GLRL	After	23,33
31	GLCM	Before	23,33
60	Wavelets	Before	23,33

Chapter 6

Conclusions and Future Work

In this project, evolutionary approaches are explored for building an evaluation function of tumor response to treatment, based on the before and after-treatment values of a set of clinical variables and image features of the lesions (extracted from PET/CT images). The need for such function is justified by two main facts: 1) manual analysis by specialists is a complex and time-consuming task, 2) some metrics for faster analysis have been proposed (e.g SUV), but with questionable reliability.

The image features were extracted from the outputs of a collection of state-of-the-art image descriptors, computed from patches of the lesions.

The whole feature set was validated in a classification experiment – it was observed that our set of features allows for very high classification performances. After that, symbolic regression, a particular application of GAs, was adopted to obtain the evaluation functions.

The preliminary results regarding the proposed approach are optimistic – an evaluation function with class-wise accuracies of 80%, 75%, 85,71% and 88,89% was obtained. However, more experiments need to be carried out regarding methods and parameters of the symbolic regression runs, so as to find the ones which optimize performance.

In addition, for generalization to be an hypothesis, the approach must be tested on larger and more balanced datasets.

Finally, we intend to extend this approach to other oncological pathologies.

Bibliography

[1] Z. Z. J. A. Siegel R1, Ma J, "Cancer statistics, 2014," *CA: Cancer J. Clin.*, pp. 9–29, 2014.

[2] W. H. Organization, *World Cancer Report 2014.* 2014.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[4] W. H. Organization, *World Cancer Report 2014.* 2014.

[5] P. E. Kinahan and J. W. Fletcher, "Pet/ct standardized uptake values (suvs) in clinical practice and assessing response to therapy," *Semin Ultrasound CT MR*, vol. 31, no. 6, pp. 496–505, 2010.

[6] R. M. Haralick, K. Shanmuga, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[7] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, Jul 2002.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[10] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[11] D. C. Moura and M. A. Guevara-López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 561–574, 2013.

[12] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. K. Jr., "The digital database for screening mammography.," in *Proceedings of the 5th international workshop on digital mammography*, pp. 212–218, 2000.

[13] M. A. Guevara-López, N. G. Posada, D. C. Moura, R. R. Pollán, and M. José, "BCDR : A BREAST CANCER DIGITAL REPOSITORY," in *15th International Conference on Experimental Mechanics*, pp. 1–5, 2015.

[14] A. Tahmasbi, F. Saki, and S. B. Shokouhi, "Classification of benign and malignant masses based on Zernike moments," *Computers in Biology and Medicine*, vol. 41, no. 8, pp. 726–735, 2011.

[15] J. Suckling, J. Parker, D. R. Dance, S. M. Astley, I. Hutt, C. R. M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. L. Kok, P. Taylor, D. Betal, and J. Savage, "The Mammographic Image Analysis Society digital mammogram database," in *Proceedings of the International Workshop on Digital Mammography*, pp. 211–221, 1994.

[16] S. Sharma and P. Khanna, "Computer-aided diagnosis of malignant mammograms using zernike moments and svm," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 77–90, 2015.

[17] M. K. Hu, "Visual-pattern recognition by moment invariants.," *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.

[18] M. R. Teague, "Image-analysis via the general-theory of moments.," *Journal of Optical Society of America*, vol. 70, no. 8, pp. 920–930, 1980.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, June 2005.

[20] B. Wu, P. Khong, and T. Chan, "Automatic detection and classification of nasopharyngeal carcinoma on PET/CT with support vector machine," *International Journal of Computer Assisted Radiology and Surgery*, vol. 7, pp. 635–646, 2012.

[21] I. E. Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. Thorstad, and J. Deasy, "Exploring feature-based approaches in {PET} images for predicting cancer treatment outcomes," *Pattern Recognition*, vol. 42, no. 6, pp. 1162 – 1171, 2009. Digital Image Processing and Pattern Recognition Techniques for the Detection of Cancer.

[22] P. M. Morgado, "Automated Diagnosis of Alzheimer' s Disease using PET Images A study of alternative procedures for feature extraction and selection Electrical and Computer Engineering," Master's thesis, MSc thesis at Electrical and Computer Engineering Dep., Higher technical institute, Technical University of Lisbon, 2012.

[23] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.," *Journal of magnetic resonance imaging : JMRI*, vol. 27, no. 4, pp. 685–691, 2008.

[24] A. Depeursinge, D. Sage, A. Hidki, A. Platon, P. Poletti, M. Unser, and H. Muller, "Lung tissue classification using wavelet frames," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007.*, pp. 6259–6262, Aug 2007.

[25] L. Boroczky, Z. Luyin, and K. Lee, "Feature subset selection for improving the performance of false positive reduction in lung nodule cad," in *Proceedings. 18th IEEE Symposium on Computer-Based Medical Systems, 2005.*, pp. 85–90, June 2005.

[26] L. Dettori and L. Semler, "A comparison of wavelet, ridgelet, and curvelet-based texture classification algorithms in computed tomography," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 486–498, 2007.

[27] A. Madabhushi, M. D. Feldman, D. N. Metaxas, J. Tomaszeweski, and D. Chute, "Automated detection of prostatic adenocarcinoma from high-resolution ex vivo mri," *IEEE Transactions on Medical Imaging*, vol. 24, no. 12, pp. 1611–1625, 2005.

[28] L. A. Meinel, A. H. Stolpen, K. S. Berbaum, L. L. Fajardo, and J. M. Reinhardt, "Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system," *Journal of Magnetic Resonance Imaging*, vol. 25, no. 1, pp. 89–95, 2007.

[29] D. Unay, A. Ekin, M. Cetin, R. Jasinschi, and A. Ercil, "Robustness of local binary patterns in brain MR image analysis.," in *IEEE Engineering in Medicine and Biology Society Annual Conference*, vol. 2007, pp. 2098–2101, 2007.

[30] K. K. Reddy, B. Solmaz, P. Yan, N. G. Avgeropoulos, D. J. Rippe, and M. Shah, "Confidence guided enhancing brain tumor segmentation in multi-parametric MRI," in *Proceedings - International Symposium on Biomedical Imaging*, pp. 366–369, 2012.

[31] P. Theocharakis, D. Glotsos, I. Kalatzis, S. Kostopoulos, P. Georgiadis, K. Sifaki, K. Tsakouridou, M. Malamas, G. Delibasis, D. Cavouras, and G. Nikiforidis, "Pattern recognition system for the discrimination of multiple sclerosis from cerebral microangiopathy lesions based on texture analysis of magnetic resonance images," *Magnetic Resonance Imaging*, vol. 27, no. 3, pp. 417 – 422, 2009.

[32] B. Verma and J. Zakos, "A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 5, pp. 46–54, March 2001.

[33] S. Halkiotis, T. Botsis, and M. Rangoussi, "Automatic detection of clustered microcalcifications in digital mammograms using mathematical morphology and neural networks," *Signal Processing*, vol. 87, no. 7, pp. 1559 – 1568, 2007.

[34] A. Papadopoulos, D. Fotiadis, and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 141 – 150, 2005.

[35] I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, "Computer aided diagnosis of breast cancer in digitized mammograms," *Computerized Medical Imaging and Graphics*, vol. 26, no. 5, pp. 309 – 319, 2002.

[36] E.-L. Chen, P.-C. Chung, C.-L. Chen, H.-M. Tsai, and C.-I. Chang, "An automatic diagnostic system for ct liver image classification," *Biomedical Engineering, IEEE Transactions on*, vol. 45, pp. 783–794, June 1998.

[37] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. Helvie, D. Adler, and M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *Medical Imaging, IEEE Transactions on*, vol. 15, pp. 598–610, Oct 1996.

[38] K. Suzuki, S. G. Armato, F. Li, S. Sone, and K. Doi, "Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography," *Medical Physics*, vol. 30, no. 7, 2003.

[39] R. Setiono, "Extracting rules from pruned neural networks for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 8, no. 1, pp. 37 – 51, 1996.

[40] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen, "Lung cancer cell identification based on artificial neural network ensembles," *Artif. Intell. Med.*, vol. 24, pp. 25–36, Jan. 2002.

[41] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," 1988.

[42] M. Talebi, "Medical ultrasound image segmentation using genetic active contour," *Journal of Biomedical Science and Engineering*, vol. 04, no. February, pp. 105–109, 2011.

[43] L. Ballerini, "Genetic Snakes for Medical Images Segmentation," *Evolutionary Image Analysis, Signal Processing and Telecommunications*, vol. 1596, pp. 59–73, 1999.

[44] Y. Lee, T. Hara, H. Fujita, S. Itoh, and T. Ishigaki, "Automated Detection of Pulmonary Nodules in Helical CT Images Based on an Improved Template-Matching Technique," vol. 20, no. 7, pp. 595–604, 2001.

[45] J. Dehmeshki, X. Ye, X. Lin, M. Valdivieso, and H. Amin, "Automated detection of lung nodules in CT images using shape-based genetic algorithm," *Computerized Medical Imaging and Graphics*, vol. 31, pp. 408–417, 2007.

[46] F. I. Corporation, Y. Sun, C. F. Babbs, and E. J. Delp, "A Comparison of Feature Selection Methods for the Detection of Breast Cancers in Mammograms : Adaptive Sequential Floating Search vs . Genetic Algorithm," pp. 6532–6535, 2005.

[47] B. Sahiner, H. P. Chan, N. Petrick, M. a. Helvie, and M. M. Goodsitt, "Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis.," *Physics in medicine and biology*, vol. 43, pp. 2853–2871, 1998.

[48] M. a. Anastasio, H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi, "A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms.," *Medical physics*, vol. 25, no. 1998, pp. 1613–1620, 1998.

[49] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, 2002.

[50] M. Tan, M. Hartley, M. Bister, and R. Deklerck, "Automated feature selection in neuroevolution," *Evolutionary Intelligence*, vol. 1, no. 4, pp. 271–292, 2009.

[51] M. Tan, R. Deklerck, B. Jansen, and J. Cornelis, "Analysis of a feature-deselective neuroevolution classifier (FD-NEAT) in a computer-aided lung nodule detection system for CT images," *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference companion - GECCO Companion '12*, p. 539, 2012.

[52] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for CT images," *Medical Physics*, vol. 38, no. 10, p. 5630, 2011.

Appendix A

Table A.1: LVQNN - Original.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	6,5	12	4	8,5	7,75
8	6,5	6	4	2,5	4,75
10	6,5	6	4	8,5	6,25
12	6,5	6	10	8,5	7,75
14	6,5	6	10	2,5	6,25
16	6,5	6	10	8,5	7,75
18	6,5	6	4	8,5	6,25
20	6,5	6	10	8,5	7,75
22	6,5	6	4	8,5	6,25
24	6,5	6	10	2,5	6,25
26	6,5	6	4	8,5	6,25
28	6,5	6	4	2,5	4,75

Table A.2: LVQNN – SMOTE.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	11,5	6	6,5	6,5	7,625
8	11,5	6	6,5	6,5	7,625
10	5,5	6	6,5	6,5	6,125
12	5,5	6	6,5	6,5	6,125
14	5,5	6	6,5	6,5	6,125
16	5,5	6	6,5	6,5	6,125
18	5,5	6	6,5	6,5	6,125
20	5,5	6	6,5	6,5	6,125
22	5,5	6	6,5	6,5	6,125
24	5,5	12	6,5	6,5	7,625
26	5,5	6	6,5	6,5	6,125
28	5,5	6	6,5	6,5	6,125

Table A.3: LVQNN – PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	6,5	12	3,5	6,5	7,125
8	6,5	6	3,5	6,5	5,625
10	6,5	6	12	11	8,875
12	6,5	6	9	6,5	7
14	6,5	6	3,5	2	4,5
16	6,5	6	3,5	11	6,75
18	6,5	6	3,5	6,5	5,625
20	6,5	6	9	6,5	7
22	6,5	6	9	2	5,875
24	6,5	6	9	11	8,125
26	6,5	6	3,5	6,5	5,625
28	6,5	6	9	2	5,875

Table A.4: LVQNN – SMOTE+PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	11,5	6,5	1,5	6	6,375
8	11,5	6,5	1,5	6	6,375
10	5,5	6,5	5,5	6	5,875
12	5,5	6,5	10,5	6	7,125
14	5,5	6,5	5,5	6	5,875
16	5,5	6,5	10,5	6	7,125
18	5,5	6,5	5,5	6	5,875
20	5,5	6,5	10,5	6	7,125
22	5,5	6,5	5,5	6	5,875
24	5,5	6,5	5,5	6	5,875
26	5,5	6,5	10,5	6	7,125
28	5,5	6,5	5,5	12	7,375

Table A.5: MLPI - Original.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	1,5	9	1	9	5,125
8	1,5	1,5	2	2	1,75
10	4,5	1,5	4	9	4,75
12	9,5	9	3	2	5,875
14	9,5	12	7,5	12	10,25
16	4,5	5,5	9,5	9	7,125
18	9,5	9	5	5	7,125
20	4,5	3,5	6	9	5,75
22	9,5	9	9,5	5	8,25
24	4,5	5,5	7,5	2	4,875
26	9,5	3,5	11,5	5	7,375
28	9,5	9	11,5	9	9,75

Table A.6: MLPI – SMOTE.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	1	1,5	6,5	1,5	2,625
8	7	11,5	1	3,5	5,75
10	7	5	10,5	1,5	6
12	7	9	3	3,5	5,625
14	7	5	10,5	7,5	7,5
16	7	5	10,5	7,5	7,5
18	7	5	6,5	5	5,875
20	7	9	3	7,5	6,625
22	7	11,5	6,5	11	9
24	7	5	3	11	6,5
26	7	9	10,5	11	9,375
28	7	1,5	6,5	7,5	5,625

Table A.7: MLPI – PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	1	1,5	6,5	1,5	2,625
8	7	11,5	1	3,5	5,75
10	7	5	10,5	1,5	6
12	7	9	3	3,5	5,625
14	7	5	10,5	7,5	7,5
16	7	5	10,5	7,5	7,5
18	7	5	6,5	5	5,875
20	7	9	3	7,5	6,625
22	7	11,5	6,5	11	9
24	7	5	3	11	6,5
26	7	9	10,5	11	9,375
28	7	1,5	6,5	7,5	5,625

Table A.8: MLPI – SMOTE + PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	8,5	11	1	3,5	6
8	8,5	3	10	1,5	5,75
10	11,5	8,5	2,5	1,5	6
12	11,5	3	2,5	3,5	5,125
14	5,5	6,5	6	5	5,75
16	8,5	11	6	6	7,875
18	2	3	10	8	5,75
20	8,5	11	12	8	9,875
22	3	6,5	10	10,5	7,5
24	5,5	8,5	6	12	8
26	4	3	6	8	5,25
28	1	3	6	10,5	5,125

Table A.9: MLPII – Original.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	2	5	3	5,5	3,875
8	8	9,5	3	1	5,375
10	8	12	10,5	2	8,125
12	8	1,5	8	10	6,875
14	2	9,5	1	10	5,625
16	8	5	8	5,5	6,625
18	8	9,5	5,5	5,5	7,125
20	8	1,5	10,5	5,5	6,375
22	8	9,5	12	5,5	8,75
24	8	5	8	10	7,75
26	2	5	3	5,5	3,875
28	8	5	5,5	12	7,625

Table A.10: MLPII -SMOTE.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	5	11,5	7	4,5	7
8	11	7	3,5	6,5	7
10	5	11,5	3,5	2	5,5
12	5	7	7	4,5	5,875
14	11	7	1	6,5	6,375
16	5	2	7	9	5,75
18	5	2	3,5	11,5	5,5
20	5	7	9	11,5	8,125
22	5	7	11	2	6,25
24	5	2	3,5	2	3,125
26	11	7	10	9	9,25
28	5	7	12	9	8,25

Table A.11: MLPII – PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	2	5	3	5,5	3,875
8	8	9,5	3	1	5,375
10	8	12	10,5	2	8,125
12	8	1,5	8	10	6,875
14	2	9,5	1	10	5,625
16	8	5	8	5,5	6,625
18	8	9,5	5,5	5,5	7,125
20	8	1,5	10,5	5,5	6,375
22	8	9,5	12	5,5	8,75
24	8	5	8	10	7,75
26	2	5	3	5,5	3,875
28	8	5	5,5	12	7,625

Table A.12: MLPII – SMOTE+PCA.

Neurons	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
6	5	6	3	1	3,75
8	9	1	1	4	3,75
10	7,5	6	6,5	4	6
12	11	2,5	10,5	10,5	8,625
14	11	6	3	4	6
16	5	6	6,5	4	5,375
18	11	9	6,5	4	7,625
20	2	11,5	6,5	7,5	6,875
22	7,5	6	10,5	10,5	8,625
24	2	10	3	7,5	5,625
26	5	2,5	10,5	10,5	7,125
28	2	11,5	10,5	10,5	8,625

Table A.13: PNN – Original.

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	2,5	13,5	15,5	15,5	11,75
1	2,5	13,5	15,5	15,5	11,75
2	2,5	13,5	15,5	15,5	11,75
4	2,5	13,5	15,5	15,5	11,75
8	5	13,5	13	12,5	11
16	11,5	6	12	12,5	10,5
32	11,5	6	11	11	9,875
64	11,5	3	8	10	8,125
128	11,5	3	8	8	7,625
256	11,5	1	8	8	7,125
512	11,5	3	8	8	7,625
1024	11,5	6	4,5	6	7
2048	11,5	8,5	4,5	5	7,375
4096	11,5	8,5	2	3,5	6,375
8192	11,5	13,5	2	1	7
16384	11,5	13,5	2	2	7,25
32768	11,5	13,5	8	3,5	9,125

Table A.14: PNN – SMOTE.

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	2,5	14	15,5	15,5	11,875
1	2,5	14	15,5	15,5	11,875
2	2,5	14	15,5	15,5	11,875
4	2,5	14	15,5	15,5	11,875
8	5	9	13	12,5	9,875
16	6	8	12	12,5	9,625
32	8	5	11	11	8,75
64	8	3,5	7,5	10	7,25
128	8	3,5	7,5	8	6,75
256	10	1,5	7,5	8	6,75
512	14	1,5	7,5	8	7,75
1024	14	6	7,5	6	8,375
2048	14	7	7,5	5	8,375
4096	14	10	2	3,5	7,375
8192	14	14	2	1	7,75
16384	14	14	2	2	8
32768	14	14	4	3,5	8,875

Table A.15: PNN – PCA

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	2,5	13,5	17	15,5	12,125
1	2,5	13,5	15,5	15,5	11,75
2	2,5	13,5	15,5	15,5	11,75
4	2,5	13,5	14	15,5	11,375
8	11	13,5	13	12,5	12,5
16	11	6	12	12,5	10,375
32	11	6	11	11	9,75
64	11	3	10	10	8,5
128	11	3	7	8	7,25
256	11	1	7	8	6,75
512	11	3	7	8	7,25
1024	11	6	7	6	7,5
2048	11	8,5	4	5	7,125
4096	11	8,5	2	3,5	6,25
8192	11	13,5	2	1	6,875
16384	11	13,5	2	2	7,125
32768	11	13,5	7	3,5	8,75

Table A.16: PNN – SMOTE+PCA.

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	1,5	14,5	17	15,5	12,125
1	1,5	14,5	15,5	15,5	11,75
2	3	14,5	15,5	15,5	12,125
4	6	10,5	14	15,5	11,5
8	8	8,5	13	12,5	10,5
16	10	8,5	11,5	12,5	10,625
32	10	3	11,5	11	8,875
64	10	3	10	10	8,25
128	6	1	9	8	6
256	6	3	8	8	6,25
512	4	5,5	6,5	8	6
1024	12	5,5	6,5	6	7,5
2048	15	7	4,5	5	7,875
4096	15	10,5	2	3,5	7,75
8192	15	14,5	2	1	8,125
16384	15	14,5	2	2	8,375
32768	15	14,5	4,5	3,5	9,375

Table A.17: RBFNN – Original

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	9	12,5	14,5	5,5	10,375
0,5	9	12,5	14,5	5,5	10,375
1	9	12,5	14,5	5,5	10,375
2	9	12,5	14,5	5,5	10,375
4	9	12,5	14,5	5,5	10,375
8	9	12,5	14,5	5,5	10,375
16	9	12,5	14,5	5,5	10,375
32	9	12,5	2	16	9,875
64	9	12,5	2	16	9,875
128	9	12,5	2	16	9,875
256	9	12,5	5	2	7,125
512	9	6,5	6	1	5,625
1024	9	1,5	7,5	12	7,5
2048	9	4	4	13,5	7,625
4096	9	1,5	7,5	9,5	6,875
8192	9	4	10	11	8,5
16384	9	4	10	13,5	9,125
32768	9	6,5	10	9,5	8,75

Table A.18: RBFNN – SMOTE.

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	12	13	14,5	11,5	12,75
s 1	12	13	14,5	11,5	12,75
2	12	13	14,5	11,5	12,75
4	12	13	14,5	11,5	12,75
8	12	13	14,5	11,5	12,75
16	12	13	14,5	11,5	12,75
32	12	13	1,5	16	10,625
64	12	13	1,5	16	10,625
128	12	13	3	16	11
256	12	6	4	2	6
512	12	6	8	1	6,75
1024	1	3	8	6	4,5
2048	3,5	8	5,5	8	6,25
4096	3,5	1	5,5	3	3,25
8192	3,5	3	8	4,5	4,75
16384	3,5	3	10,5	7	6
32768	6	6	10,5	4,5	6,75

Table A.19: RBFNN – PCA

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	9	12	8,5	3	8,125
1	9	12	1	15	9,25
2	9	2,5	17	3	7,875
4	9	5	13,5	15	10,625
8	9	1	15	15	10
16	9	12	16	3	10
32	9	12	2,5	15	9,625
64	9	12	2,5	15	9,625
128	9	12	13,5	3	9,375
256	9	12	12	3	9
512	9	12	4,5	12	9,375
1024	9	12	8,5	10,5	10
2048	9	12	11	7	9,75
4096	9	12	10	10,5	10,375
8192	9	5	4,5	9	6,875
16384	9	2,5	6,5	8	6,5
32768	9	5	6,5	6	6,625

Table A.20: RBFNN – SMOTE+PCA.

Spread	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
0,5	11,5	11	11	10	10,875
1	11,5	11	10	6	9,625
2	11,5	11	1	14	9,375
4	11,5	11	2,5	14	9,75
8	11,5	11	15	2,5	10
16	11,5	11	17	2,5	10,5
32	11,5	11	4	14	10,125
64	11,5	11	2,5	14	9,75
128	11,5	11	16	2,5	10,25
256	11,5	11	14	2,5	9,75
512	11,5	11	13	5	10,125
1024	11,5	11	6,5	9	9,5
2048	5	11	12	14	10,5
4096	1	1,5	9	14	6,375
8192	2,5	1,5	8	14	6,5
16384	2,5	4	6,5	8	5,25
32768	4	3	5	7	4,75

Table A.21: kNN – Original.

k	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
1	15,5	1,5	26	29,5	18,125
2	15,5	1,5	26	29,5	18,125
3	15,5	3,5	28	27	18,5
4	15,5	6,5	29,5	27	19,625
5	15,5	3,5	29,5	13	15,375
6	15,5	19,5	26	13	18,5
7	15,5	6,5	11,5	13	11,625
8	15,5	6,5	5,5	27	13,625
9	15,5	6,5	18,5	13	13,375
10	15,5	19,5	18,5	13	16,625
11	15,5	19,5	18,5	13	16,625
12	15,5	19,5	11,5	13	14,875
13	15,5	19,5	18,5	13	16,625
14	15,5	19,5	18,5	13	16,625
15	15,5	19,5	18,5	13	16,625
16	15,5	19,5	18,5	13	16,625
17	15,5	19,5	18,5	13	16,625
18	15,5	19,5	18,5	13	16,625
19	15,5	19,5	18,5	13	16,625
20	15,5	19,5	18,5	13	16,625
21	15,5	19,5	18,5	13	16,625
22	15,5	19,5	5,5	13	13,375
23	15,5	19,5	5,5	13	13,375
24	15,5	19,5	5,5	13	13,375
25	15,5	19,5	5,5	13	13,375
26	15,5	19,5	5,5	13	13,375
27	15,5	19,5	5,5	13	13,375
28	15,5	19,5	5,5	13	13,375
29	15,5	19,5	5,5	13	13,375
30	15,5	19,5	5,5	13	13,375

Table A.22: kNN – SMOTE.

k	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
1	11,5	1,5	16	29,5	14,625
2	11,5	1,5	16	29,5	14,625
3	11,5	3	28	27	17,375
4	11,5	7,5	29,5	27	18,875
5	3,5	7,5	29,5	13	13,375
6	11,5	7,5	16	13	12
7	11,5	13,5	9	13	11,75
8	11,5	7,5	16	27	15,5
9	3,5	13,5	24	13	13,5
10	11,5	16,5	16	13	14,25
11	11,5	16,5	16	13	14,25
12	11,5	23,5	9	13	14,25
13	11,5	23,5	24	13	18
14	3,5	23,5	16	13	14
15	1	23,5	24	13	15,375
16	3,5	30	24	13	17,625
17	11,5	23,5	24	13	18
18	24	7,5	24	13	17,125
19	24	7,5	24	13	17,125
20	24	7,5	16	13	15,125
21	24	7,5	9	13	13,375
22	24	13,5	9	13	14,875
23	24	13,5	16	13	16,625
24	24	23,5	9	13	17,375
25	24	23,5	4,5	13	16,25
26	24	23,5	4,5	13	16,25
27	24	23,5	4,5	13	16,25
28	24	23,5	4,5	13	16,25
29	24	23,5	1,5	13	15,5
30	24	23,5	1,5	13	15,5

Table A.23: kNN – PCA.

k	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
1	15,5	1,5	26,5	29,5	18,25
2	15,5	1,5	26,5	29,5	18,25
3	15,5	3	29,5	27	18,75
4	15,5	6	29,5	27	19,5
5	15,5	6	28	13	15,625
6	15,5	19,5	25	13	18,25
7	15,5	6	11,5	13	11,5
8	15,5	6	5,5	27	13,5
9	15,5	6	18,5	13	13,25
10	15,5	19,5	18,5	13	16,625
11	15,5	19,5	18,5	13	16,625
12	15,5	19,5	11,5	13	14,875
13	15,5	19,5	18,5	13	16,625
14	15,5	19,5	18,5	13	16,625
15	15,5	19,5	18,5	13	16,625
16	15,5	19,5	18,5	13	16,625
17	15,5	19,5	18,5	13	16,625
18	15,5	19,5	18,5	13	16,625
19	15,5	19,5	18,5	13	16,625
20	15,5	19,5	18,5	13	16,625
21	15,5	19,5	18,5	13	16,625
22	15,5	19,5	5,5	13	13,375
23	15,5	19,5	5,5	13	13,375
24	15,5	19,5	5,5	13	13,375
25	15,5	19,5	5,5	13	13,375
26	15,5	19,5	5,5	13	13,375
27	15,5	19,5	5,5	13	13,375
28	15,5	19,5	5,5	13	13,375
29	15,5	19,5	5,5	13	13,375
30	15,5	19,5	5,5	13	13,375

Table A.24: kNN – SMOTE + PCA.

k	Negative	Neutral	Positive (partial)	Positive (complete)	Average rank
1	14,5	2	13,5	29,5	14,875
2	14,5	2	13,5	29,5	14,875
3	10	2	30	27	17,25
4	10	4,5	23,5	27	16,25
5	12	4,5	23,5	13	13,25
6	10	6	13,5	13	10,625
7	5,5	23,5	13,5	13	13,875
8	5,5	11	13,5	27	14,25
9	5,5	11	18,5	13	12
10	5,5	11	13,5	13	10,75
11	5,5	11	23,5	13	13,25
12	5,5	11	18,5	13	12
13	1,5	11	23,5	13	12,25
14	1,5	11	23,5	13	12,25
15	14,5	11	23,5	13	15,5
16	14,5	11	23,5	13	15,5
17	23,5	18,5	28,5	13	20,875
18	23,5	18,5	28,5	13	20,875
19	23,5	18,5	23,5	13	19,625
20	23,5	18,5	13,5	13	17,125
21	23,5	18,5	9	13	16
22	23,5	23,5	13,5	13	18,375
23	23,5	18,5	5	13	15
24	23,5	23,5	5	13	16,25
25	23,5	23,5	5	13	16,25
26	23,5	28	5	13	17,375
27	23,5	28	5	13	17,375
28	23,5	28	5	13	17,375
29	23,5	28	5	13	17,375
30	23,5	28	1	13	16,375