POLITECNICO DI TORINO

Master of Science in Biomedical Engineering

Master's Degree Thesis

Characterization of Prostate Cancer aggressiveness based on bi-parametric MRI.

Supervisors

Prof. Samanta ROSATI

Prof. Valentina GIANNINI

Candidate

Giulia NICOLETTI

December 2020

Abstract

The aim of this study is to provide a noninvasive, radiological image-based Computer Aided Diagnosis (CAD) able to distinguish between high aggressive (Gleason Score (GS) >= 4+3) and low-aggressive (GS<= 3+4) Prostate Cancers (PCa). The system exploits the use of Machine Learning (ML) and Deep Learning (DL) on biparametric Magnetic Resonance Images coming from Candiolo IRCCS and San Giovanni Molinette hospital.

Regarding the ML approach, once tumor areas have been manually segmented, features of first order statistics, intensity-based and texture features are extracted, both from T2WI and ADC maps. The study carries out a parallel analysis of ten different Datasets, which differ in type of feature (3D or 2D), voxel spacing, application of filters, and bin number.
Datasets have been pre-processed using some data cleaning techniques, then Univariate Analysis and Multi-parametric Analysis are carried out. The Univariate Analysis involves the calculation of the area under the ROC curves (AUC) of each feature, Mann Witney U test, and correlation analysis, both between each feature vector and the output (classification). The Multi-parametric analysis includes the Genetic Algorithm (GA), the Minimum Redundancy Maximum Relevance (MRMR), and the Affinity Propagation (AP) methods. Four Feature Selection strategies have been carried out: the first one consists of evaluating the 7-fold cross-validation performances of the model trained with an increasing number of features, added one by one in descending order of AUC, until the overfitting point is found; the others use the subsets resulting from the three multivariable algorithms. At the end, the best ML classifier is a svm, that achieves excellent performance in the training set (100% accuracy), good results in the test set (75%, 70% and 85%, respectively of accuracy, sensitivity, and specificity) and slightly lower results in the validation set (64%, 56%, and 100% of accuracy, sensitivity, and specificity respectively).

Regarding the DL approach, once the ROIs (3x3 and 5x5 pixel, totally inside the lesion) have been extracted, both from T2WI and ADC maps, Convolutional Neural Networks (CNN) with 1, 2, and 3 Convolutional Layers are tested. Several CNNs are trained, different in size and number of filters, number of neurons, and set parameters. The resulting best DL classifier achieves good performance in the training set (71%, 72%, and 71% respectively of accuracy, sensitivity, and specificity), low performance in the test set (44%, 30% and 67% , respectively of accuracy, sensitivity, and specificity) and slightly higher results in the validation

set (82%, 94%, 25% respectively of accuracy, sensitivity, and specificity).

The results from ML and DL approaches show lower results in the validation sets due to the low ability of the classifiers to generalize the problem. In particular, the best ML model achieves better performance than the best DL one. The generalization problem must be reduced increasing the number of samples in the datasets and also reformulating the division of patients into training, testing, and validation sets, in order to obtain a training set that is more representative of the variability of the two tumor classes.

Table of Contents

Chapter 1

Introduction

1.1 Prostate Cancer (PCa)

1.1.1 Prostate gland anatomy

The prostate gland is situated in the true pelvis and it is present only in men. Its main function is to secrete one of the components of semen, the prostate fluid. The muscles of the prostate gland also help propelling this seminal fluid into the urethra during ejaculation. The prostate is located between the penis and the bladder, and surrounds the urethra (see figure 1.1).

Gross anatomy - The prostate gland is divided into five lobes: anterior and posterior lobes, two lateral lobes, and one median lobe. In clinics, it is described as having two lateral lobes, right and left, and a median lobe [2].

Microanatomy - The prostate gland is composed of histologically different zones; based on these differences, the gland is divided into three anatomical zones [3].

1. The **Peripheral Zone** (PZ) is the largest zone ($\sim 70\%$) of the gland. It surrounds most of the central zone and partially surrounds the distal part of the prostatic urethra. Most PCas ($\sim 70 - 80\%$ [2, 4]) are located in the peripheral zone and may be detected by Digital Rectal Examination (DRE) when the volume is > 0.2 mL. In $\sim 18\%$ of cases, PCa is detected by suspect DRE alone, irrespective of PSA level [5, 6].

2. The **Central Zone** (CZ) is a cone-shaped region that forms the base of the gland and surrounds the ejaculatory ducts. It covers 25% of glandular tissue in young adults [4]. The incidence rate of PCa in this area is low ($\sim$ 2,5% [4]), but the tumors appear to be particularly aggressive and tend to spread to the seminal vesicles.

3. The **Transition Zone** (TZ) is a small glandular zone (in young men, accounts

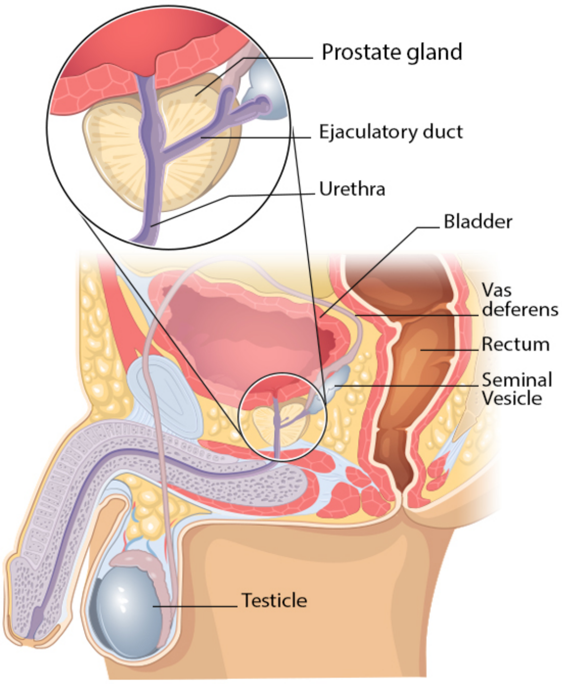


Figure 1.1: Prostate anatomy [1].

for only 5-10% of prostatic glandular tissue [4]). It surrounds a portion of the urethra between the urinary bladder and verumontanum. The incidence of PCa is $\sim$ 20% [4]. As it increases in size, it is the area responsible for Benign Prostatic Hyperplasia (BPH).

The lower part of the prostate, called the Apex, is surrounded by muscle and fibrous tissue. This area is known as the Fibromuscular Stroma. Finally, the capsule (a fibrous layer) surrounds the entire prostate [2].

1.1.2 PCa epidemiology

Prostate cancer is the most frequently diagnosed cancer in men in 12 regions of the world [8]. In the United States, it accounts alone for more than 1 in 5 new diagnoses [9]. In terms of new cases, it is the first leading type of cancer in Africa, Americas and Europe [8].

In 2020 in America, about 191,930 new cases and about 33,330 deaths from PCa are estimated [10]. In 2020 in the EU, it is the third predicted cause of cancer men deaths, with 78 800 deaths and a rate of 10.0/100 000 [11].

Figure 1.3 shows the mortality trend in six EU countries, starting from the

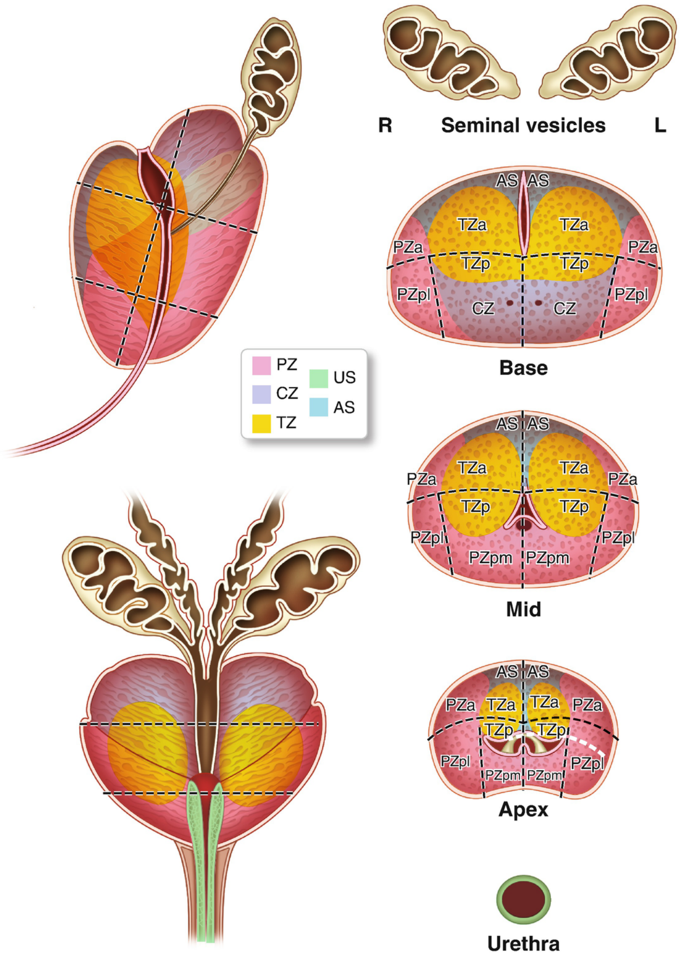


Figure 1.2: Prostate zones (PZ: Peripheral Zones, CZ: Central Zone, TZ: Transition Zones, US: Urethral Sphincter, AS: Anterior fibromuscular Stroma, a: anterior, p: posterior, mp: medial posterior, lp: lateral posterior) [7].

1970s up to the predictions for 2020. There are four main trends: the curves of Italy and France increase until 1990 and then begin to decrease; a similar trend is reported in UK and Germany with a temporarily shift of about 5 years; Spain's curve increases until 2000; lastly, Poland has the worst trend, as it increases until 2000s, decreases slightly and has a growing predicted trend for 2020.

Table 1.1 shows the mortality rate of PCa, starting from the year 2005 up to the prediction for 2020, for the same six EU countries discussed above. Between them, Italy has the best predictions for 2020 in terms of Age-Standardized Mortality rate

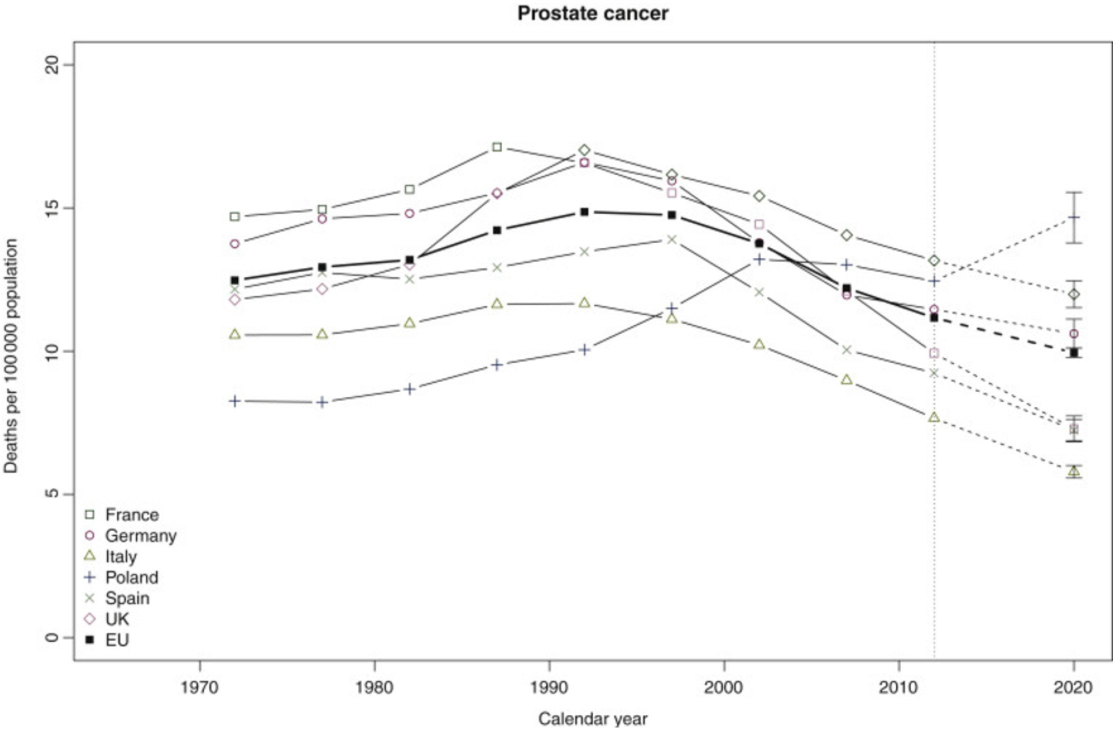


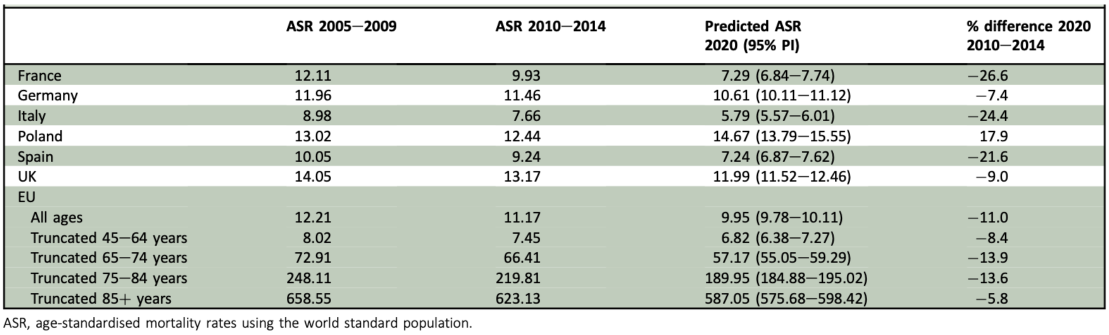
Figure 1.3: Age-standardised (world population) cancer mortality rate trends for all ages in quinquenniums from 1970–1974 to 2010–2014 and predicted rates for 2020 with 95% prediction intervals for prostate cancer in studied countries and in the EU as a whole. [11]

(ASR), while on the opposite side there is Poland, that has the worst prediction for 2020.

The differences between these six European countries remain largely unexplained, some may be related to the differences between the treatments used for PCa in each country [11].

1.1.3 PCa aetiology

A systematic review of autopsy studies reported a prevalence of PCa at age <30 years of 5%, to a prevalence of 59% by age >79 years [6]. Figure 1.4 shows the mortality rate of PCa for different age groups, starting from 1970 up to the predictions of 2020. It is immediately evident how the increase in age affects the mortality rate, remaining one of the most significant risk factors of the PCa [6].



	ASR 2005−2009	ASR 2010−2014	Predicted ASR 2020 (95% PI)	% difference 2020 2010−2014
France	12.11	9.93	7.29 (6.84−7.74)	−26.6
Germany	11.96	11.46	10.61 (10.11−11.12)	−7.4
Italy	8.98	7.66	5.79 (5.57−6.01)	−24.4
Poland	13.02	12.44	14.67 (13.79−15.55)	17.9
Spain	10.05	9.24	7.24 (6.87−7.62)	−21.6
UK	14.05	13.17	11.99 (11.52−12.46)	−9.0
EU				
All ages	12.21	11.17	9.95 (9.78−10.11)	−11.0
Truncated 45−64 years	8.02	7.45	6.82 (6.38−7.27)	−8.4
Truncated 65−74 years	72.91	66.41	57.17 (55.05−59.29)	−13.9
Truncated 75−84 years	248.11	219.81	189.95 (184.88−195.02)	−13.6
Truncated 85+ years	658.55	623.13	587.05 (575.68−598.42)	−5.8

ASR, age-standardised mortality rates using the world standard population.

Table 1.1: Age-standardised prostate cancer mortality rates for all ages in selected European countries and for the EU as a whole for all ages [11].

PCa also appears to be affected by genetic predisposition: the probability of contracting this type of tumor increases in men with father or brothers who already had it. However, hereditary PCa appears to affect the age of onset (the tumor occurs six to seven years earlier) but not the aggression and the clinical course. [6]

Furthermore, there are a large number of environmental/exogenous factors associated with the risk of developing PCa or progressing from latent to clinical PCa.

- Obesity, associated with a lower risk of low-grade PCa, but increased risk of high-grade PCa [6].
- Dietary factors, e.g. high alcohol intake, but also total abstention from alcohol has been associated with a higher risk of PCa and PCa-specific mortality [6].
- Cigarette smoking, associated with an increased risk of PCa death [6].
- Sexually transmitted infections, e.g. Gonorrhoea, were significantly associated with an increased incidence of PCa.

However, there are currently no specific preventive or dietary measures recommended to reduce the risk of developing PCa.

1.1.4 PCa diagnosis

PCa screening remains a debated topic today. What is needed is to break the direct link between diagnosis and active treatment [6], in order to avoid over-treatments. The diagnosis is traditionally performed by monitoring the Prostate-Specific Antigen (PSA) level, the level of a protein produced by prostate cells in man's blood. This type of tumor causes an increment of PSA level, but being PSA organ and not

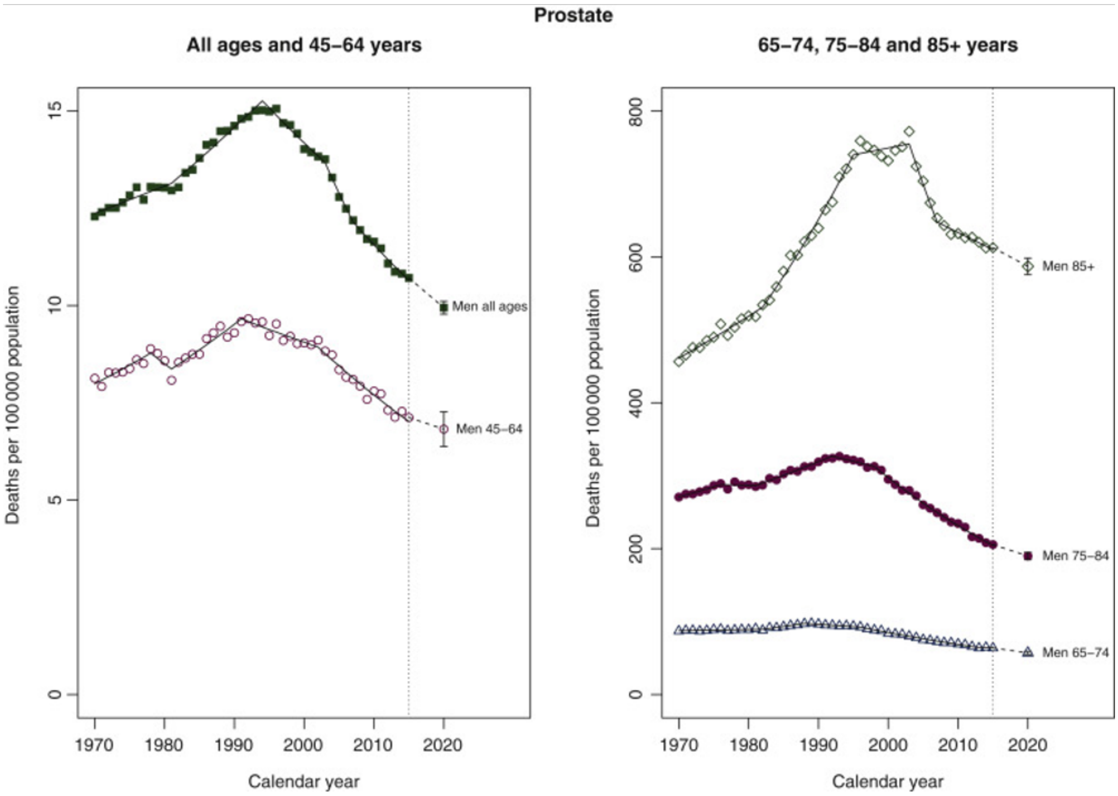
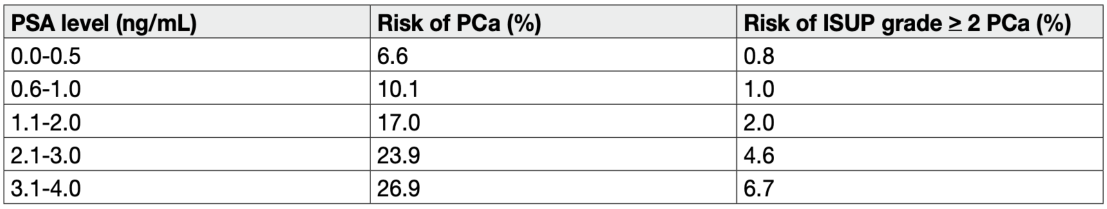


Figure 1.4: Annual prostate cancer age-standardized (world population) death rates in the EU per 100 000 for all ages, 45-64, 65-74, 75-84, and 85+ age-groups from 1970 to 2015, the resulting joinpoint regression models, and predicted rates for the year 2020 with 95% prediction intervals. On the left all ages (full squares) and 45-64 (empty circles) age groups; on the right 65-74 (empty triangles), 75-84 (full circles), and 85+ (empty diamonds) age groups. [11].

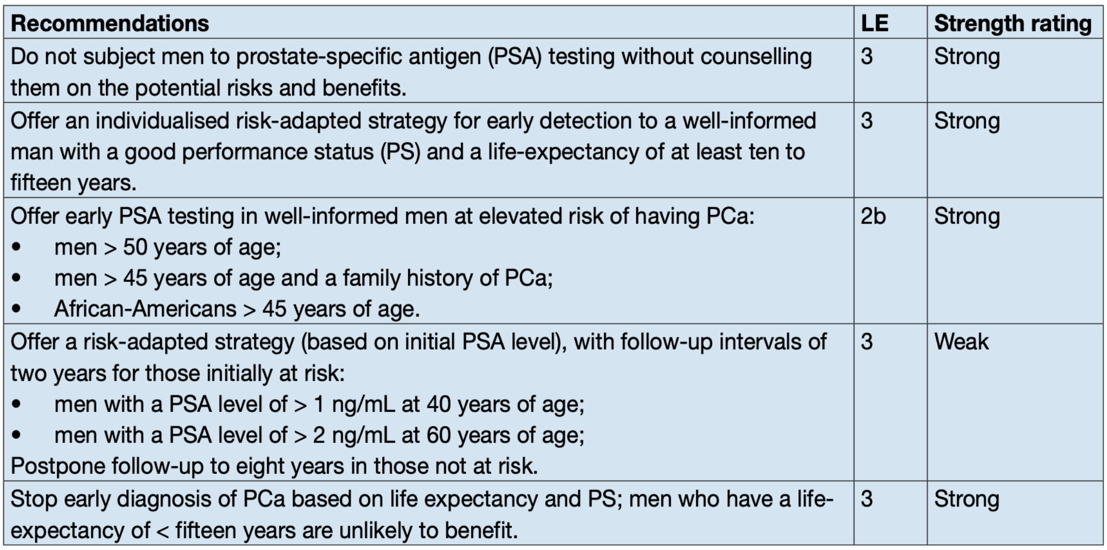
cancer specific [6], it may be the symptom of non-malignant conditions, such as Benign Prostatic Hypertrophy (BPH) or prostatitis. Furthermore, a PSA level lower than 4.0 ng/mL was previously considered normal but over time it has been seen that even low values can be associated with PCa. Table 1.2 shows the risk of clinically significant PCa incidence in relation to low PSA levels. Table 1.3 shows the official urology guidelines for screening and early detection.

After the PSA test, the Digital Rectal Examination (DRE) is usually done. This test allows detecting PZ PCa when the lesion volume is $>$ 0.2 mL [6]. Based on the suspicious outcome of the PSA and DRE tests, the definitive diagnosis depends on histopathological verification of adenocarcinoma in prostate biopsy cores or specimens from the TransUrethral Resection of the Prostate (TURP) or



PSA level (ng/mL)	Risk of PCa (%)	Risk of ISUP grade ≥ 2 PCa (%)
0.0-0.5	6.6	0.8
0.6-1.0	10.1	1.0
1.1-2.0	17.0	2.0
2.1-3.0	23.9	4.6
3.1-4.0	26.9	6.7

Table 1.2: Risk of PCa in relation to low PSA values [6].



Recommendations	LE	Strength rating
Do not subject men to prostate-specific antigen (PSA) testing without counselling them on the potential risks and benefits.	3	Strong
Offer an individualised risk-adapted strategy for early detection to a well-informed man with a good performance status (PS) and a life-expectancy of at least ten to fifteen years.	3	Strong
Offer early PSA testing in well-informed men at elevated risk of having PCa: • men > 50 years of age; • men > 45 years of age and a family history of PCa; • African-Americans > 45 years of age.	2b	Strong
Offer a risk-adapted strategy (based on initial PSA level), with follow-up intervals of two years for those initially at risk: • men with a PSA level of > 1 ng/mL at 40 years of age; • men with a PSA level of > 2 ng/mL at 60 years of age; Postpone follow-up to eight years in those not at risk.	3	Weak
Stop early diagnosis of PCa based on life expectancy and PS; men who have a life-expectancy of < fifteen years are unlikely to benefit.	3	Strong

Table 1.3: Guidelines for screening and early detection [6].

prostatectomy for Benign Prostatic Enlargement (BPE) [6].

1.1.5 PCa grading

Traditionally, PCa grading is done using the Gleason Score (GS), a scale of 1 to 5 referring to normal cells and tumor cells, respectively. The pathologist looking at the biopsy sample assigns one Gleason grade to the most predominant pattern and a second Gleason grade to the second most predominant pattern. The two grades are then added together to determine the final score.

Based on the GS, PCa is divided into three risk bands that refer to the probability of tumor progression: low risk (GS 3+3), intermediate-risk (GS 3+4 or 4+3), and high risk (GS $\geq$ 4+4). In making decisions about PCa screening, diagnosis, and treatment, not only the patient's age but also the patient's life expectancy, health status, and comorbidities must be considered [6]. Considering these factors, generally radical treatment is done in high-risk PCa, while active treatment of intermediate-risk PCa is uncertain. Some 3+4 PCa with low volume may be candidates for active surveillance. Moreover, Gleason 7 prostate cancer shows heterogeneous behavior, conferring to GS 3+4 and GS 4+3 different specific prognosis [12] and mortality [13, 14, 15].

In this regard, to further define the clinically highly significant distinction between GS 3+4 and 4+3 [16], the 2014 ISUP endorsed grading system [17] limits the number of PCa grades, ranging them from 1 to 5: 1 (GS 2-6), 2 (GS 3+4), 3 (GS 4+3), 4 (GS 8), 5 (GS 9-10). This approach has the potential to reduce overtreatment, at the same time allowing a more accurate grade stratification than previous systems. Moreover, starting from 1 and not 6, it also helps to reduce fear among patients.

1.2 MRI for Prostate Cancer

It is evident that the spread of PSA screening of healthy men has allowed a decrease in the PCa mortality rate, but, on the other hand, it has led to a consequent increase in the number of diagnoses and treatment of many clinically insignificant lesions. For this reason, it is fundamental to be able to distinguish between clinically significant and non-clinically significant lesions, and among the clinically significant ones to understand which are the aggressive ones, that need active treatment, and which ones can be followed with active surveillance.

In this regard, MRI is the imaging modality of choice for the PCa local staging.

1.2.1 Multi-parametric and bi-parametric MRI

Multiparametric MRI (mpMRI) generally consists of three imaging sequences.

- **Diffusion-Weighted Imaging** (DWI) measures the mobility of water molecules due to Brownian motion. From DWI the **Apparent Diffusion Coefficient** (ADC) is calculated. Specifically, PCa results in an increase in cell density, so it is detected in the areas with the highest signal (white) on DWI and, instead, in the areas with a lower signal (black) in the corresponding ADC map.

- **T2-Weigthed Imaging** (T2WI) reflects local tissue water. It is the sequence that outlines the anatomy of the prostate. In this image, the transitional and peripheral areas are clearly delineated: the first is represented as a high signal area (white), while the second as an area with both high and low signals. In T2WI, PCa delineates zones of moderately low signal intensity.

- **Dynamic Contrast-Enhanced** (DCE) images are obtained after the injection of a gadolinium contrast medium. This imaging sequence shows the local vascular environment and, therefore, allows to recognize areas with altered vascularity, typical of PCa.

Due to the different sequence procedures, obtaining a mpMRI takes a long time. Furthermore, the use of a contrast agent and an endorectal coil are invasive procedures for the patient, with several undesirable effects. To overcome these drawbacks, biparametric MRI (bpMRI) is considered to be an alternative for the detection [18]. Unlike mpMRI, it is composed only of T2W and DW sequences.

Nowadays, bpMRI is becoming increasingly useful in the detection and characterization of PCa.

1.2.2 MRI based CADx

The evolution in imaging techniques of the last decades has allowed improvements in the detection and characterization of tumors. The interpretation of these images, however, remains dependent on various factors such as the experience of the observing clinician and the complexity of the pathology. In this regard, it has been demonstrated how assisted detection and diagnosis technologies can help and improve the clinician's work. Some researchers have shown how the detection performance of PCa by less experienced observers increases, thanks to this kind of system, up to the same performance as more experienced observers, at the same time increasing the reader agreement [19] [20]. Moreover, it helps experienced observers to recognize more CS lesion patients while decreasing the overall reading time [21]. These technologies, which do not replace the doctor's decision, serve as a second opinion and are known by the acronym CAD. CADs include two sub-branches: Computer Aided Detection (CADe), system that helps in tumor detection, and Computer Aided Diagnosis (CADx), which helps in its characterization [22].

MRI is a particularly useful tool in the creation of CAD systems. Over the past 20 years, a large number of mpMRI-based CAD systems have been implemented, based on both machine learning and deep learning techniques. R. R. Wildeboer et al. [23] presented a detailed overview of the published CAD designs applied to PCa. What is interesting to underline is that of the 83 researches cited, 71 deal with CAD systems based on MR images. Furthermore, most of these studies address the problem of PCa detection, creating CADe systems. Among the studies that, on the other hand, focus on the characterization of PCa and its grading (CADx), many researchers tend to classify between clinically significant (CS) (GS ≥ 7) and not clinically significant GS ≤ 6) lesions.

As mentioned in section 1.1.5, it would be useful to focus on the differentiation of intermediate-risk tumors, given their heterogeneity, creating classifiers capable

of recognizing lesions with GS $\leq 3+4$ from those with GS $\geq 4+3$. Currently, few studies have dealt with this type of classification. R. Cao et al. [24] implemented a multi-class CNN to jointly detect PCa lesions and predict their GS groups, starting from mp-MRI, and obtained an AUC value of 0.81 and 0.79 for the classifications of CS PCa (GS $\geq 3+4$) and PCa with GS $\geq 4+3$, respectively.
D. Fehr et al. [25] proposed an svm model, based on first- and second-order texture features from T2W images and ADC maps and sample augmentation through oversampling techniques to compensate the unbalanced dataset, to both classify between CS vs not-CS PCas and between 3+4 vs 4+3 GS lesions. In the first case, they reached a 93% accuracy for cancers occurring in both PZ and TZ and 92% for cancers occurring in the PZ alone; in the second case, they obtained a 92% and a 93% accuracy for PZ-TZ and PZ alone, respectively. They also tried to create the same classifiers starting from the ADC maps alone, obtaining an accuracy of about 60%, demonstrating, in this way, how the combination of ADC maps with T2W sequences helps in the classification of the PCa.
P. Tiwari et al. [26] implemented an ensemble classifier, called Semi-Supervised Multi Kernel Graph Embedding, starting from T2W and MRS, to classify between benign versus cancerous, and high (GS $\geq 4+3$)) versus low (GS$\leq 3+4$) Gleason grade reaching an AUC value of 0.89 and 0.84, respectively.
C. Fusun et al. [27] evaluated linear discriminant analysis (LDA) and support vector machine (SVM) classifiers to predict Gleason Groups (GG), using age, the presence of a palpable prostate abnormality, PSA level, index lesion size, and Likert scales of T2W, DW, and DCE, as features. They reached mean sensitivities of 86.51% and 87.88% and mean specificities of 63.99% and 56.83% for LDA and SVM, respectively.

Therefore, considering the need to distinguish intermediate grade PCa and the potential of bpMRI, the aim of this study is to develop a bi-parametric based CADx system which can automatically distinguish between low-aggressive GS $\leq$ 3+4 and high-aggressive GS $>$ 3+4 lesions, classifying lesions in 0 and 1, respectively.

Chapter 2

Machine Learning for Prostate Cancer Aggressiveness characterization

2.1 Dataset

2.1.1 Patients

The present study involves the processing of biparametric MR images of patients from Candiolo IRCCS and San Giovanni Molinette hospital. Inclusion in the study requires the fulfillment of the following requirements:

- Biopsy confirmed PCa;
- Bi-parametric MR without endorectal coil and without contrast medium.

Specifically, some lesions are excluded from the study, their quantity and motivation are shown in the flowchart in figure 2.1.

2.1.2 Features

In the case of machine learning techniques, the model creation phase is preceded by feature extraction and feature selection phases. There are basically two ways to extract features: voxel-wise or region-wise [22]. Specifically, in the first case it can be intensity-, edge-, texture- or position- based; in the second, statistical-, histogram- or anatomical- based.

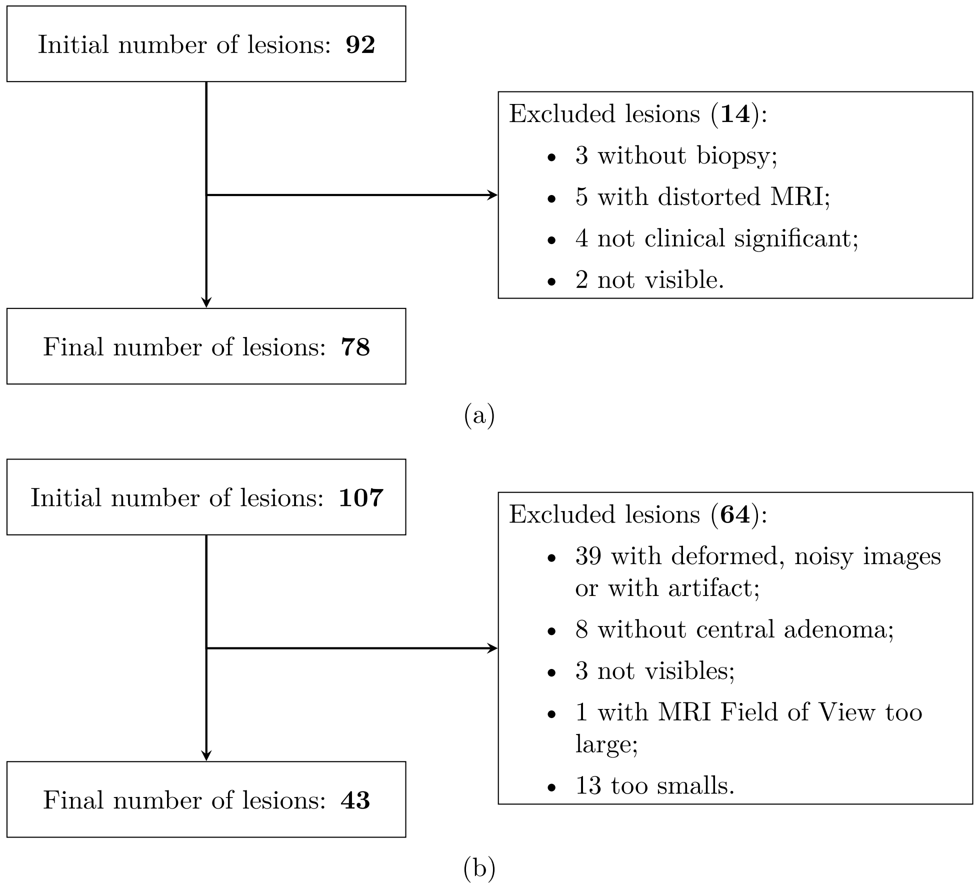

Initial number of lesions: **92**
Excluded lesions (**14**):
• 3 without biopsy;
• 5 with distorted MRI;
• 4 not clinical significant;
• 2 not visible.
Final number of lesions: **78**
(a)
Initial number of lesions: **107**
Excluded lesions (**64**):
• 39 with deformed, noisy images or with artifact;
• 8 without central adenoma;
• 3 not visibles;
• 1 with MRI Field of View too large;
• 13 too smalls.
Final number of lesions: **43**
(b)

Figure 2.1: Flowchart of lesions coming from Candiolo IRCCS (a) and San Giovanni Molinette hospital (b).

Several studies assessed the association between Haralick texture features and PCa aggressiveness [28, 29] extracted from T2W and ADC images.
T.W. Baek et al. [30] analyzed the correlation coefficient between GS and texture features (first-order statistics and second-order statistics based on the gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), and wavelet transformation features) and applied a multiple regression to the significant parameters demonstrating the association between GS and GLCM entropy.
C. Jensen et al. [31] used Image histogram and texture features to create a k-nearest neighbor (kNN) classifier to distinguish lesions into their Grade Group, obtaining AUC values equal to 0.88 and 0.96 in PZ lesions, 0.89, 0.83 in TZ lesions, in the distinction between GS 3+4 vs others and GS $\leq 3 + 4$ vs others, respectively.
A. Chaddad et al. [32] analyzed the combination between Joint Intensity Matrix (JIM) and GLCM for predicting PCa GS, obtaining AUC values of 78.40% for

GS≤6, 82.35% for GS 3+4, and 64.76% for GS $\geq$ 4+3.

Regarding this study, texture, intensity-based and histogram features are extracted from bi-parametric MRI. Specifically, to avoid loss of reproducibility and validation, feature are extracted using an in-house software compliant with the Image Biomarker Standardization Initiative (IBSI), implemented using C ++ and ITK libraries. Texture features are extracted from:

- The **Grey Level Co-occurrence Matrix** (GLCM), that describes how many times a certain gray value is close to another gray value, along a certain direction. Their proximity is defined as inter-pixel distance and is a free parameter chosen by the user (in this case, set equal to 1 pixel).

- The **Grey-Level Run Length Matrix** (GLRLM) describes the image in a directional way by counting the *runs*, or series of pixels, where the same gray value is found consecutively along a certain direction.

Given a 26-connected neighborhood in the 3D case and an 8-connected neighborhood in the 2D one, and an inter-pixel distance equal to 1, the possible directions on which to calculate the GLCM and the GLRM are 13 and 4, respectively. Once the matrices have been calculated along all the directions, they are averaged, and the image descriptors are calculated on them (see table 2.3 (a) and (b)). In addition, the volume of the ROI in mm3 is used as parameter. In the case of the ADC dataset, also the mean Intensity of the Histogram, kurtosis, and Intensity-based statistical features (table 2.3 (c)) are added. Table 2.1 summarizes the number and the type of features extracted for each dataset.

All these features are extracted both three-dimensional (3D) and two-dimensional (2D). Specifically, in the 3D domain texture features need isotropic voxel spacing in order to be rotationally invariant. For this reason, first a pre-interpolation filter (Gaussian, $\sigma = 0.5mm$) is applied to the image and then voxel spacing is downsampled from $0.31 \times 0.31 \times 0.5mm^3$ to $0.5 \times 0.5 \times 0.5mm^3$. In the 2D domain, features are extracted both from $0.31 \times 0.31mm^2$ and $0.5 \times 0.5mm^2$ pixels, and both with and without the application of the Gaussian filter ($\sigma = 0.5mm$). All the feature extraction configurations described so far are repeated with a fixed number of bins equal to 32 and 64. A total of twenty datasets are created, ten extracted from T2W images and ten from ADC maps (see table 2.2).

	GLCM	**GLRM**	**intensity-based**	**others**	**total**
ADC	25	16	15	3	59
T2	25	16	-	1	42

Table 2.1: Number and type of extracted features for each dataset.

	Voxel spacing (mm)	**Type of feature**	**Filter**	**Number of Bins**
1	0.5	3D	Gaussian ($\sigma = 0.5$ mm)	32bin
2	0.5	3D	Gaussian ($\sigma = 0.5$ mm)	64bin
3	0.5	2D	Gaussian ($\sigma = 0.5$ mm)	32bin
4	0.5	2D	Gaussian ($\sigma = 0.5$ mm)	64bin
5	0.5	2D	No Blur	32bin
6	0.5	2D	No Blur	64bin
7	0.31	2D	Gaussian ($\sigma = 0.5$ mm)	32bin
8	0.31	2D	Gaussian ($\sigma = 0.5$ mm)	64bin
9	0.31	2D	No Blur	32bin
10	0.31	2D	No Blur	64bin

Table 2.2: Type of datasets created by extrapolating features both from ADC maps and T2W images.

a) GLCM

1. Joint max
2. Joint Average
3. Joint Variance
4. Joint Entropy
5. Difference Average
6. Difference Variance
7. Difference Entropy
8. Sum Average
9. Sum Variance
10. Sum Entropy
11. Angular Second Moment
12. Contrast
13. Dissimilarity
14. Inverse Difference
15. Normalised Inverse Difference
16. Inverse Difference Moment
17. Normalised Inverse Difference
18. Inverse Variance
19. Correlation
20. Autocorrelation
21. Cluster Tendency
22. Cluster Shade
23. Cluster Prominence
24. Information Measure of Correlation I
25. Information Measure of Correlation II

b) GLRM

26. Short Run Emphasis (SRE)
27. Long Run Emphasis (LRE)
28. Low Grey level Run Emphasis (LGRE)
29. High Grey level Run Emphasis (HGRE)
30. Short Run Low Grey level Emphasis (SRLGE)
31. Short Run High Grey level Emphasis (SRHGE)
32. Long Run Low Grey level Emphasis (LRLGE)
33. Long Run High Grey level Emphasis (LRHGE)
34. Grey Level Non-Uniformity (GLNU)
35. Normalised Grey Level Non-Uniformity (GLNU-norm)
36. Run Length Non-Uniformity (RLNU)
37. Normalised Run Length Non-Uniformity (RLNU-norm)
38. Run Percentage (RP)
39. Grey Level Variance (GL-var)
40. Run Length Variance (RL-var)
41. Run Entropy (RE)

c) intensity-based

43. Mean Intensity
44. Minimum Intensity
45. Maximum Intensity
46. Intensity Range
47. 1 Intensity Percentile
48. 10 Intensity Percentile
49. 25 Intensity Percentile
50. 50 Intensity Percentile
51. 75 Intensity Percentile
52. 90 Intensity Percentile
53. 95 Intensity Percentile
54. Intensity Interquartile Range (IQR)
55. Intensity Skewness
56. Intensity Kurtosis
57. Intensity Variance

Table 2.3: Features.

2.2 Univariate analysis

Univariate feature analysis evaluates the strength of the relationship that each feature has, individually, with the outcome. This is useful to better understand the different datasets and therefore the characteristics of the data. Thus, in this section, we will analyze all feature vectors independently of each other, coming from the 20 datasets.

2.2.1 AUC

The ability of each feature to correctly classify high and low aggressive lesions is evaluated. To do this, the **Area Under the Curve** (AUC) is calculated, i.e. the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed by modifying the cut-off value beyond which the lesion is considered cancerous (Positive) and below which it is considered non-cancerous (Negative). For each cut-off value, the following calculations are made:

- number of True Positive (TP), i.e. number of positive lesions correctly classified;
- number of True Negative (TN), i.e. number of negative lesions correctly classified;
- number of False Positive (FP), i.e. number of negative lesions classified as positive;
- number of False Negative (FN), i.e. numbero of positive lesions classified as negative.

At this point, two performance indices are calculated:

- Specificity, $\frac{TN}{TN+FP}$
- Sensitivity, $\frac{TP}{TP+FN}$

The ROC curve is drawn using 1-Specificity (false positive fraction) on the x-axis and Sensitivity (true positive fraction) on the y-axis as coordinates for each tested cut-off value. The aim is to find the cut-off that maximizes the value of the ordinate axis and minimizes that of the abscissa, i.e. the cut-off able to classify the positive lesions as correctly as possible without mistaking the negative ones. An AUC value of 1 corresponds to a cut-off able to discriminate between the positive and negative lesions with 100% sensitivity and 100% specificity.
For our purposes, the AUC values greater than or equal to 0.7 are considered to be good performances.

The frequency with which each lesion is misclassified is therefore analyzed: only the features with AUC value greater than 0.7 are considered and those patients who are incorrectly classified by all the features are highlighted, patients that, potentially, can never be classified correctly.

2.2.2 Statistical test

A statistical test is applied to understand if there are features that are closer to the output (real lesion classification) than others. It is a nonparametric test, which means that no assumption is made about the distribution of the population. Given

$$X_{(1)}, X_{(2)}, ..., X_{(m)} \text{ and } Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$$

two samples representing respectively the feature vector and the output, the alternate null hypothesis is that the two samples come from the same population, that is

$$H_0 : F_Y(x) = F_X(x) \text{ for all } x$$

where F_X and F_Y are the two populations of the two samples X and Y, respectively. The test is performed between each feature and the output with a significance level of the decision (α) equal to 0.05.

Mann-Whitney U-test

The **Mann-Whitney (M-W) U-test** is the non-parametric analog of Student's t-test for independent samples. It tests the equality of population medians of X and Y.

The assumption made on the data is that the two samples are drawn from continuous distributions, so that the possibility $X_i = Y_j$ for some i and j need not be considered. The M-W U-test statistic, U, is defined as the number of times a y precedes an x in an ordered arrangement of the elements, in the two independent random samples X and Y [33]. U is the smaller of U_X and U_Y, defined as below:

$$U_X = n_X n_Y + \tfrac{n_Y(n_Y+1)}{2} - R_X,$$

$$U_Y = n_X n_Y + \tfrac{n_X(n_X+1)}{2} - R_Y$$

where n_X and n_Y are the sizes of X and Y, R_X and R_Y are the sum of the ranks in X and Y (the sum of the ranks of a sample is defined as the sum of the positions that the data of that sample occupies after being sorted from smallest to largest).

The theoretical range of U is from 0 (complete separation between the two samples, H_0 most likely false) to $n_X * n_Y$ (little evidence in support of H1). Note that $U_X + U_Y = n_X n_Y$ is always true.

2.2.3 Correlation feature-output

The correlation coefficient measures the linear dependence of X and Y. It is defined as

$$\rho(X,Y) = \tfrac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where $cov(X,Y)$ is the covariance of X and Y, and σ_X and σ_Y are the standard deviation of X and Y, respectively.
A correlation coefficient close to 1 means that the feature is strongly linearly correlated to the outcome, vice versa a value close to 0 indicates lack of linear dependence. The sign of the coefficient states whether the dependence is positive (as one variable increases, the other also increases) or negative (as one increases, the other decreases). For our purpose, it is important to evaluate only the magnitude of the correlation coefficient and not its sign. So, an absolute value greater than or equal to 0.7 is considered to be a good performance.

2.3 Multiparametric analysis

Multiparametric analysis is a type of feature selection that evaluates subsets of features, taking into account not only the predictive power of the features but also the correlation between them. Ideally, the aim is to obtain a feature subset that keeps all the informative content of the dataset and reduces the redundancy by eliminating highly correlated and uninformative features.
This analysis is carried out on the ADC and T2 datasets, and also on a dataset including all the ADC and T2 features (hereafter called *ADC-T2 joined* dataset).

2.3.1 Dataset division

The division of the dataset into training and test set is a critical phase: the aim is to obtain a training set that is sufficiently large and representative of the entire dataset. For the multiparametric study, only the lesions coming from Candiolo IRCCS are used, as the only ones available at the time of the analysis. Specifically, lesions are sorted in ascending order of volume and divided into three equally numerous volume bands, as shown in figure 2.2. Thus, 4 large, 4 medium, and 2 small lesions are chosen randomly, and the corresponding patients, with all their lesions, are selected to constitute the test set. The remaining patients are included in the training set. All the lesions coming from San Giovanni Molinette hospital are left out to create an external validation set. Figure 2.3 shows the distribution of the size of the lesions according to the three bands mentioned above, and the distribution of the two classes (0 = low-aggressive, 1 = high-aggressive) in training and test sets.

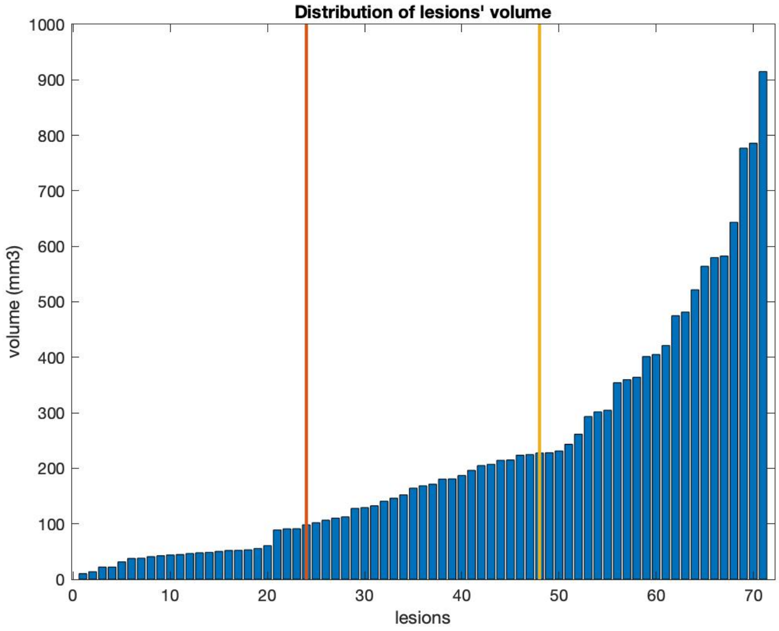


Figure 2.2: Lesions' volume distribution.

2.3.2 Minimum Redundancy Maximum Relevance

As the name suggests, the **Minimum Redundance Maximum Relevance** (MRMR) method looks for a subset composed of features that are minimally related to each other but that are maximally predictive of the outcome. It exploits the concept of Mutual Information (MI) by using the mutual information quotient (MIQ) value.

Specifically, considering two variables X and Z, the MI between them is defined as

$$I(X,Z) = \sum_{i,j} P(X = x_i, Z = z_j) log \frac{P(X=x_i, Z=z_j)}{P(X=x_i)P(Z=z_j)}.$$

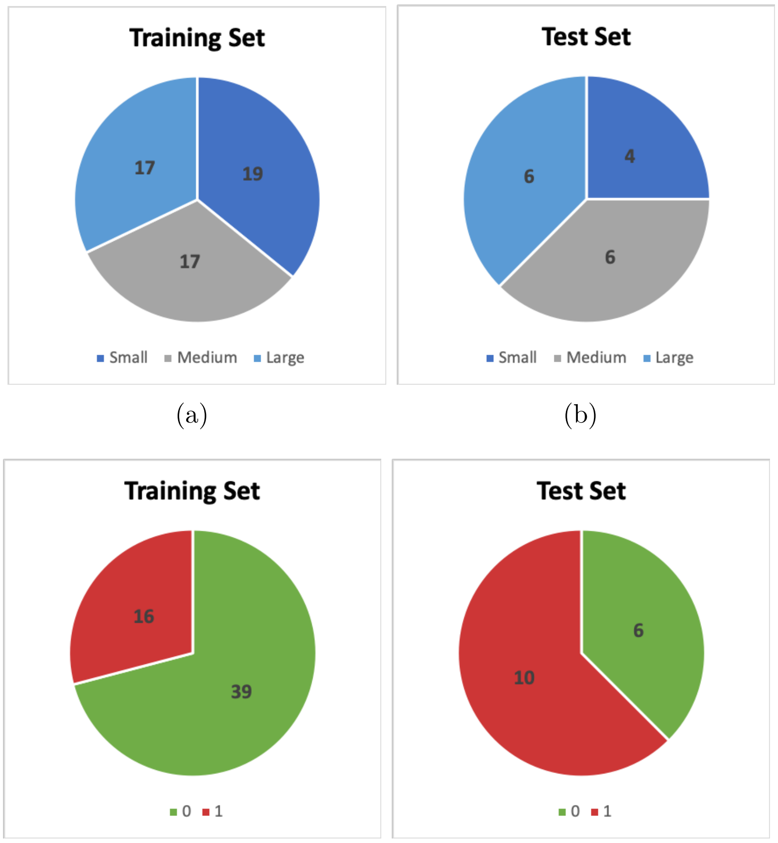
In particular, I(X,Z) will be equal to 0 if X and Z are independent of each other, on the contrary if X and Z are the same variable then I(X,Z) will correspond to its entropy.

Given the feature x, its MIQ value is defined as the ratio between relevance and redundancy of x, as follow

$$MIQ_x = \frac{V_x}{W_x}$$

with

$$V_x = I(x,y) \text{ and } W_x = \frac{1}{|S|} \sum_{z \in S} I(x,z)$$



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

Figure 2.3: Distribution of lesion volume in training set (a) and test set (b), and distribution of the two classes in training set (c) and test set (d).

where y is the outcome and |S| is the number of features in the optimal subset (S). The Matlab function used, *fsmrmr*, ranks all features in descending order based on their MIQ value. It assign a score to each feature, as high as the importance of that feature. Furthermore, if the features are ordered according to the score, the greater the confidence with which a feature is chosen as important, the lower the score of the next feature.

In this way the optimal subset will be composed of the first N features with the highest MIQ. In particular, in our case N features will be selected whose score is greater than 10^{-}2.

2.3.3 Genetic Algorithm

The **Genetic Algorithm** (GA) is a search heuristic that applies a probabilistic approach to feature selection. The name refers to the concept of Charles Darwin's theory of natural evolution, whereby the strongest individual survives and genetic mutations that are beneficial to an individual's survival are passed on through reproduction.

Specifically, each solution of the algorithm, called *Chromosome*, is encoded as a sequence of characters representing the selected feature subset. Between coding methods, the most popular is the binary one, where 0 means feature not selected and 1 feature selected. The goodness of each chromosome is evaluated through an objective function, called *fitness*, that is the ability of an individual to compete with other individuals and describes also how well that solution has adapted to the considered problem. A set of chromosomes constitutes a *Population*.

Starting from a first randomly generated population, new ones of the same size as the first are generated, iteration by iteration. The new populations are formed starting from the previous one by applying Genetic Operators:

- **Selection**: it takes the chromosomes with the highest fitness value (strongest individuals) from the previous population and inserts them into the new population. Then, a random fitness threshold is set: chromosomes with a value greater than the threshold are selected as parents of the next generation.
- **Crossover**: it cuts two parents in a random crossover point and joins the two halves, forming two children.
- **Mutation**: it changes one or more parent bits.

In this way the new population will consist of the best individuals of the previous population plus their children, generated using crossover and mutation operators. The algorithm continues to generate new populations until a stop criterion is reached: when the maximum number of iterations is reached or the fitness value no longer changes.

The implementation of GA involves several choices: the number of genes in each chromosome, the number of individuals in each population, the number of iterations and repetitions, the probability of mutation and crossover, and a fitness formulation, specific for the problem considered.

The number of genes in each chromosome is usually equal to the number of selectable features (1 bit = 1 feature). However, it is possible to add some bits to codify the value of a certain parameter, that in this way can be optimized, without setting it a priori.

Specifically, in our case, the gene number changes according to the type of fitness used.

- **Svm-based fitness**: it involves the use of a svm classifier, which is trained with the training set and with the subset of features selected by the chromosome taken in consideration and whose quality is to be evaluated. Therefore, the fitness value depends on the performance of the trained model applied to the test set.

- **MIQ-based fitness**: it is based on the concept of MIQ, explained in section 2.3.2. This fitness formulation does not depend on the performances of a classifier but only on the predictive power of the selected features.

In the case of svm-based fitness, the classifier is tested both with polynomial and Gaussian kernels. In addition, the trained svm model returns for each lesion the probability of belonging to one of the two classes, 0 (GS $<= 3 + 4$) and 1 (GS$> = 4 + 3$). Usually, lesions are classified on the basis of their probability of belonging to class 1: if it is greater than 50% then the lesion will be classified as belonging to class 1, on the contrary to class 0. However, sometimes the use of a threshold different from 50% results in better classification performances. For this reason, we have decided to test a fixed threshold, different from 50%, and also to encode the value of this classification threshold in the five final bits of the chromosome, in order to be automatically optimized by the algorithm. So, the number of genes will be equal to the feature number of the considered dataset (59 in the ADC dataset, 42 in the T2 dataset, and 100 in the ADC-T2 joined dataset) in the case with the fixed threshold, and equal to the feature number plus five bits in the case with optimized threshold.
Tables 2.4 and 2.6 show the number of genes used and the fitness formulations tested, respectively. Note that in any case the aim is to minimize the value of fitness, as it is formulated as one minus the performance of the subset of features (which must be maximized). Lastly, table 2.5 shows the other parameters set for the GA.

Dataset	**Number of features**	**Number of Genes**		
		MIQ-based	Svm-based	
			50% threshold	optimized threshold
ADC	59	59	59	64
T2	42	42	42	47
ADC-T2 joined	100	100	100	105

Table 2.4: Number of genes set for MIQ- and svm- based fitness for the different datasets and for the two types of probability threshold chosen for the svm classification.

Parameter	**Value**
number of individuals	500
number of iterations	2500
number of parents	80% of #individuals
number of repetitions	5
mutation probability (pm)	20%
crossover probability (pc)	100%

Table 2.5: Parameters set for the GA.

	fitness
svm-based	$1 - \frac{\text{sensitivity+specificity}}{2}$ $1 - \frac{\text{sensitivity+specificity}}{2} + \frac{\text{n° of selected features}}{\text{total n° of features}}$ $1 - \text{f1score}$
MIQ-based	$\frac{1}{\lvert S \rvert} \sum_{x \in S} \text{MIQ}_x$

Table 2.6: Fitness formulations tested.

2.3.4 Affinity Propagation

Affinity Propagation (AP) is a clustering algorithm, which, unlike others, does not require to specify the number of clusters a priori. It is based on the concept of "message passing" between elements.

The algorithm associates each feature with an exemplary one: all elements with the same specimen constitute a cluster. In simple terms, each element sends a message to all the others informing about its affinity towards them. In turn, the other elements respond to senders informing about their association availability. This exchange of messages continues until each element is associated with a single exemplar. In the end, the feature selection is made by taking only the exemplars as the optimal feature subset.

The algorithm consist in creating four matrices:

1. **Similarity Matrix** (s), it measures the similarity between each pair of elements, according to a certain measure of similarity. The result is a symmetric matrix with zeroes in the diagonal. These zeroes must be replaced with a

value on which the number of clusters obtained will depend: the lower the value, the lower the number of clusters. For this study, among various methods available, Mutual Information is chosen as similarity metric. In addition, the diagonal value is set equal to the minimum value present in *s*.

2. **Responsability Matrix** (r), it describes how much the k-th element is suitable as exemplar of the i-th element. The value of each cell is calculated with the following formula:

   $$r(i,k) = s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}$$

3. **Availability Matrix** (a), it describes how much the k-th element is available to be the exemplar of the i-th element. The value of each cell of the diagonal is calculated with the following formula:

   $$a(k,k) = \sum \max_{i' \neq k}\{0, r(i',k)\}$$

   While, the value of other cells, except that of the diagonal, is calculated with the following formula:

   $$a(i,k) = min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} max\{0, r(i',k)\}$$

4. **Criterion Matrix** (c), is obtained summing the availability matrix and responsibility matrix.

The last step of the algorithm is the selection of the highest value from each row of the matrix c: this cell will constitute the *exemplar* of that element (row). All elements with the same exemplar constitute a cluster.

2.4 Classifier construction

Once the features have been extracted and selected, the creation of the model involves a training phase and a testing phase.

2.4.1 Dataset division

For the construction of the classifiers, the division of Candiolo lesions carried out for the multiparametric analysis is used (see section 2.3.1). In addition, 21 Molinette lesions are included in the training set and the remaining 22 are left out, as an external validation set. Pie charts in figure 2.4 show the distribution of the two classes in training (a), test (b), and validation (c) set.

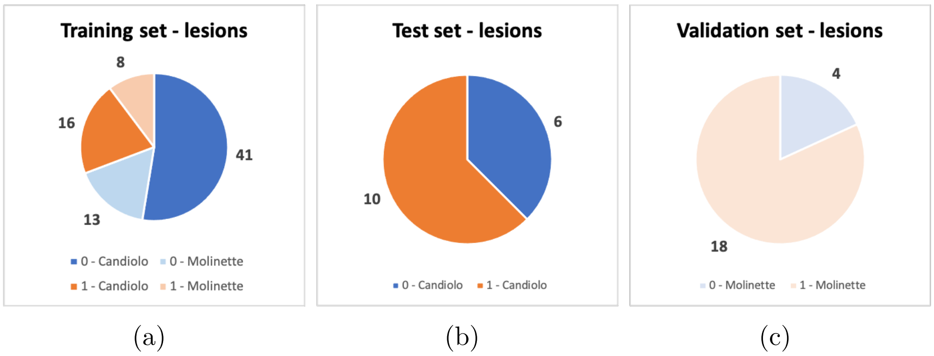


Figure 2.4: Distribution of the two classes in training (a), test (b), and validation (c) set.

2.4.2 Models

The trained classifiers can be categorized according to the Feature Selection method performed. Specifically, the FS techniques that are carried out are five: two deriving from the Univariate Analysis (described in section 2.2), and the remaining three from Multiparametric Analysis (described in section 2.3).

1st FS method: AUC ranking (only features with AUC>70%)

The first FS method implemented involves the use of an increasing number of features to train the model, by adding them one at a time in order of decreasing AUC value. K-fold crossvalidation is used to train and test the model, and, since the dataset is made up of about seventy lesions, a k equal to seven is used in order to obtain folds containing about ten lesions. The classifier used is the Support Vector Machine (svm) with polynomial kernel, and only the features with AUC greater than 70% are used for each of the twenty datasets, described in section 2.1.2. This analysis aims to find the number of feature corresponding to the overfitting point, i.e. the maximum number of features beyond which the classifier no longer learns from the training data. To get a general idea of the performance of the classifier, the values of accuracy, sensitivity and specificity of the training phases (with 6/7 fold) are averaged. The classification of the test set (seventh fold) is put into a vector at each iteration. At the end of the seven iterations the vector contains the prediction of all the lesions used as a test and so the test performance indices are calculated.

2nd FS method: AUC ranking (all features)

The feature subset is chosen according to AUC values, in the same way as in the first FS method. However, this time the performance of the models are evaluated until all the features are added, and not only those with AUC>70%. The trained classifiers are svm with polynomial and Gaussian kernel, and Random Forest (RF) classifier with 100 trees.
In addition, the classification threshold is not set to 50% but it is optimized: the best cut-off value that allows the best compromise between sensitivity and specificity is chosen. Thus, lesions with probability of belonging to class 1 (high-aggressive tumor) greater than this threshold will be classified in class 1, otherwise in class 0 (low-aggressive lesion).
Furthermore, the correlation between pair of features is taken into account: for each pair of features with correlation greater than or equal to a set threshold, the feature with the lowest AUC is deleted. The correlation thresholds tested are 0.99, 0.98 and 0.95.

3rd FS method: MRMR

Features selected by MRMR algorithm are used for the creation of the svm classifiers, with polynomial and Gaussian kernel. In this case, only one correlation threshold at 0.99 is set.

4th FS method: GA

Features subsets obtained with GA algorithm are used to train svm models with polynomial and Gaussian kernel.

5th FS method: AP

Lastly, exemplar features selected by Affinity Propagation are used to create svm classifiers with polynomial and Gaussian kernel.

2.5 Results

2.5.1 Univariate analysis

AUC

The obtained AUC values for the features coming from ADC maps and T2W images are shown in table 2.7 and 2.8, respectively. Features with AUC value greater than 70% are highlighted in bold and their amount in each dataset is shown in figure 2.5. Note that the numbering of the x axis corresponds to the different dataset configurations described in table 2.2. From the bar diagram it can be seen that 3D datasets (corrisponding to number 1 and 2 on the x axis) created from T2W images (red bars) are those with the greatest number of features with AUC>70%.

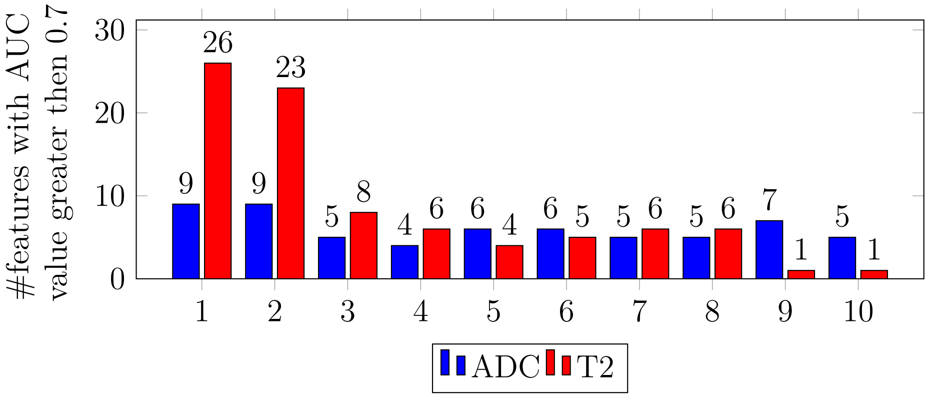


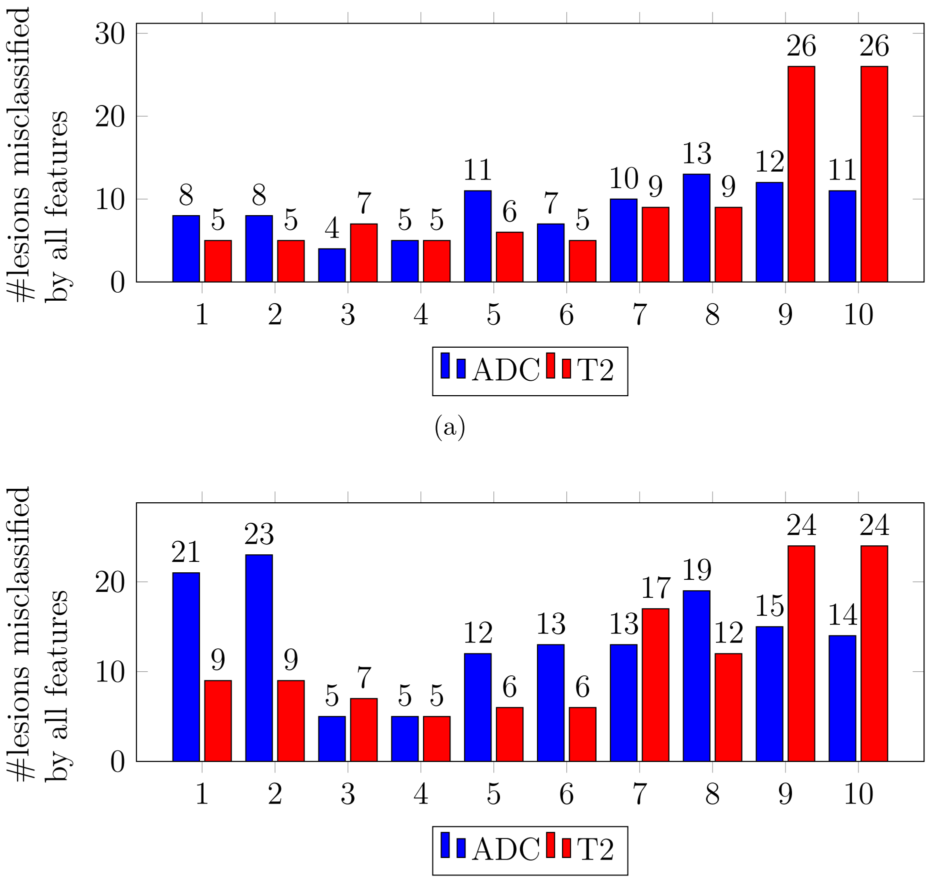
Figure 2.5: Number of features with AUC value greater then 70% coming from ADC maps and T2W images, for each of the 10 datasets (see table 2.2).

Results of the analysis of patients misclassified by all features with AUC>70% is shown in figure 2.6 (a): the bar diagram shows that all datasets have at least four lesions misclassified from all features. However, since the number of features with AUC greater than 70% varies between datasets, and since the greater the number of features used the greater the probability that at least one feature correctly classifies a given lesion, the first 4 features are selected in descending order of AUC from each dataset, and the analysis of misclassified lesions is repeated (2.6 (b)). Moving from bar diagram (a) to (b), it can be seen that:

- Both in the 32 and 64 bin cases of the ADC 3D datasets with gaussian filter and interpolation at 0.5 mm (blue bars (1) and (2)), the removal of five features causes an increase of 13 and 15 misclassified lesions, respectively, while for the T2 datasets (red bars (1) and (2)), the removal of 22 and 19 features, respectively, causes an increase of only 4 misclassified lesions.

- In the remaining cases, from (3) to (8), the number of features with AUC value greater than 0.7 is always close to 4 so there are no large changes in the number of misclassified lesions.

- The two worst cases are that of the T2 no blur dataset with original spacing (0.31 mm) 32 and 64 bin (red bars (9) and (10)) which have only one feature with AUC value greater than 0.7, and the increase from 1 to 4 features (from (a) to (b)) leads to a decrease in the misclassified lesions number of only 2, remaining the two cases with the higher number of lesions not classified correctly.



(a)

(b)

Figure 2.6: Number of lesions misclassified by all features in the 10 datasets (see table 2.2): considering all features with AUC value > 0.7 (a) and considering only the first 4 features selected in descending order of AUC (b).

Statistical test

The results in terms of p-value of the Mann-Whitney U test for the datasets from ADC maps and T2W images are shown in the table 2.9 and 2.10, respectively. Features with p-value less than 0.05 (statistically significant) are highlighted in bold. It can be seen that datasets coming from T2 images have a greater number of significant parameters than those from ADC images, despite this also the latter have a large number of significant variables. Therefore, the results of this analysis do not lead to obvious differences between the ADC and T2 datasets. Moreover, if compared to the other datasets, 3D ones, both ADC and T2, contain a greater number of features that are able to discriminate the two classes in a statistically

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	**0,741**	**0,741**	0,698	0,698	0,699	0,699	0,692	0,692	0,692	0,692
mean_ROI_STAT	0,564	0,564	0,563	0,563	0,565	0,565	0,564	0,564	0,563	0,563
minimum_ROI_STAT	0,603	0,603	0,612	0,612	0,614	0,614	0,607	0,607	0,609	0,609
maximum_ROI_STAT	**0,513**	**0,513**	**0,518**	**0,518**	0,518	0,518	0,524	0,524	0,530	0,530
Range_STAT	**0,707**	**0,707**	**0,712**	**0,712**	0,692	0,692	0,691	0,691	0,675	0,675
1stPercentile_STAT	0,601	0,601	0,598	0,598	0,605	0,605	0,599	0,599	0,603	0,603
10thPercentile_STAT	0,597	0,597	0,582	0,582	0,588	0,588	0,577	0,577	0,581	0,581
25thPercentile_STAT	0,588	0,588	0,578	0,578	0,576	0,576	0,578	0,578	0,577	0,577
50thPercentile_STAT	0,582	0,582	0,573	0,573	0,573	0,573	0,571	0,571	0,571	0,571
75thPercentile_STAT	0,554	0,554	0,554	0,554	0,559	0,559	0,553	0,553	0,557	0,557
90thPercentile_STAT	0,536	0,536	0,547	0,547	0,548	0,548	0,461	0,461	0,462	0,462
95thPercentile_STAT	0,502	0,502	0,505	0,505	0,508	0,508	0,508	0,508	0,511	0,511
IQR_STAT	0,647	0,647	0,621	0,621	0,629	0,629	0,614	0,614	0,609	0,609
skewness_STAT	0,663	0,663	0,556	0,556	0,543	0,543	0,588	0,588	0,581	0,581
kurtosis_STAT	0,566	0,566	0,562	0,562	0,556	0,556	0,609	0,609	0,620	0,620
IntensityKurtosis_STAT	0,567	0,567	0,554	0,554	0,554	0,554	0,607	0,607	0,618	0,618
IntensityVariance_STAT	0,667	0,667	0,641	0,641	0,627	0,627	0,638	0,638	0,629	0,629
mean_intensity_IH	0,632	0,632	0,532	0,535	0,522	0,521	0,545	0,547	0,554	0,554
JointMax_GLCM	0,599	0,497	0,592	0,599	0,539	0,606	0,641	0,501	0,613	0,512
JointAverage_GLCM	0,615	0,615	0,529	0,529	0,520	0,520	0,546	0,547	0,547	0,545
JointVariance_GLCM	0,565	0,565	0,583	0,582	0,607	0,606	0,599	0,597	0,650	0,649
JointEntropy_GLCM	0,658	0,534	0,543	0,629	0,557	0,648	0,628	0,574	0,608	0,574
diffAverage_GLCM	0,690	0,689	0,647	0,649	0,650	0,652	0,655	0,659	0,660	0,660
diffVariance_GLCM	0,676	0,676	0,602	0,606	0,592	0,590	0,613	0,614	0,626	0,627
diffEntropy_GLCM	0,686	0,686	0,635	0,629	0,627	0,629	0,646	0,655	0,639	0,647
sumAverage_GLCM	0,615	0,615	0,529	0,529	0,520	0,520	0,546	0,547	0,547	0,545
sumVariance_GLCM	0,532	0,535	0,559	0,557	0,576	0,571	0,588	0,591	0,631	0,629
sumEntropy_GLCM	0,565	0,552	0,495	0,550	0,495	0,575	0,562	0,502	0,574	0,530
angularSecondMoment_GLCM	0,664	0,568	0,546	0,604	0,501	0,616	0,686	0,531	0,655	0,501
contrast_GLCM	0,681	0,683	0,631	0,630	0,627	0,628	0,650	0,650	0,644	0,646
dissimilarity_GLCM	0,690	0,689	0,647	0,649	0,650	0,652	0,655	0,659	0,660	0,660
InverseDifference_GLCM	0,693	0,696	0,666	0,672	0,686	0,683	0,682	0,692	0,699	0,700
NormalisedInverseDifference_GLCM	0,690	0,691	0,652	0,653	0,659	**0,658**	0,661	**0,661**	**0,665**	**0,663**
InverseDifferenceMoment_GLCM	0,692	0,697	0,670	0,677	0,678	**0,702**	0,681	**0,702**	**0,700**	**0,713**
NormalisedInverseDifferenceMoment_GLCM	0,684	0,686	0,633	**0,630**	0,629	0,628	0,649	0,651	0,649	**0,648**
inverseVariance_GLCM	0,691	0,697	0,665	**0,708**	0,641	0,697	0,628	0,682	0,659	**0,717**
correlation_GLCM	0,655	0,654	0,612	0,612	0,612	0,612	0,601	0,601	0,600	0,603
Autocorrelation_GLCM	0,607	0,607	0,535	0,535	0,535	0,534	0,562	0,562	0,563	0,563
clustertendency_GLCM	0,532	0,535	0,559	0,557	0,576	0,571	0,588	0,591	0,631	0,629
clustershad_GLCMe	0,651	0,653	0,555	0,553	0,544	0,544	0,547	0,541	0,531	0,529
clusterprominence_GLCM	0,477	0,474	0,547	0,550	0,582	0,583	0,563	0,569	0,604	0,604
infCorr1_GLCM	0,655	0,567	0,521	0,647	0,568	0,678	0,592	0,573	0,545	0,621
infCorr2_GLCM	**0,648**	**0,570**	0,466	0,610	**0,575**	0,651	0,565	0,567	**0,537**	0,611
SRE_GLRLM	**0,708**	**0,702**	**0,666**	0,624	**0,720**	**0,694**	**0,683**	**0,675**	**0,707**	**0,665**
LRE_GLRLM	**0,730**	**0,717**	**0,706**	0,669	**0,745**	**0,723**	**0,735**	**0,726**	**0,747**	**0,715**
LGRE_GLRLM	0,538	0,548	0,611	0,644	0,650	0,697	0,618	0,652	0,605	0,629
HGRE_GLRLM	0,631	0,636	0,570	0,570	0,564	**0,564**	0,566	0,568	0,568	0,570
SRLGE_GLRLM	0,549	0,559	0,617	0,646	0,659	**0,710**	0,635	0,647	0,624	0,643
SRHGE_GLRLM	0,639	0,644	0,574	0,573	0,570	0,559	0,581	0,578	0,596	0,584
LRLGE_GLRLM	0,503	0,530	0,590	0,640	0,606	0,665	0,547	0,610	0,561	0,592
LRHGE_GLRLM	**0,562**	**0,606**	**0,491**	**0,531**	**0,518**	**0,524**	**0,538**	**0,509**	**0,545**	**0,522**
GLNU_GLRLM	**0,735**	**0,732**	**0,704**	**0,703**	**0,721**	**0,718**	**0,706**	**0,702**	**0,700**	**0,706**
GLNU_norm_GLRLM	**0,538**	**0,526**	**0,561**	**0,587**	0,496	0,595	**0,553**	0,529	0,576	0,519
RLNU_GLRLM	**0,746**	**0,744**	**0,702**	**0,702**	**0,693**	0,700	**0,703**	0,695	**0,699**	0,695
RLNU_norm_GLRLM	**0,706**	**0,700**	0,664	0,614	**0,721**	**0,696**	**0,682**	**0,665**	**0,700**	0,666
RP_GLRLM	**0,717**	**0,709**	0,691	0,641	**0,739**	**0,712**	**0,710**	**0,716**	**0,735**	0,692
GreylevelVariance_GLRLM	**0,563**	**0,562**	**0,592**	0,594	**0,635**	**0,648**	**0,594**	**0,605**	**0,635**	**0,641**
RunlengthVariance_GLRLM	**0,741**	**0,725**	**0,725**	0,685	**0,738**	**0,728**	**0,750**	**0,738**	**0,745**	**0,720**
RunEntropy_GLRLM	0,665	0,656	0,645	0,645	0,658	0,656	0,655	0,654	0,683	0,630

Table 2.7: AUC values of features derived from ADC maps. In the rows the features and in the columns the datasets. Values greater than 0.7 are highlighted in bold.

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	**0,740**	**0,740**	0,697	0,697	0,699	0,699	0,693	0,693	0,692	0,692
JointMax_GLCM	0,622	0,602	0,557	0,501	0,598	0,622	0,575	0,510	0,506	0,594
JointAverage_GLCM	0,649	0,647	0,659	0,658	0,625	0,625	0,624	0,624	0,572	0,572
JointVariance_GLCM	**0,744**	**0,741**	**0,703**	**0,702**	0,690	0,693	**0,730**	**0,730**	0,674	0,675
JointEntropy_GLCM	**0,702**	0,635	0,543	0,581	0,546	0,610	0,618	0,507	0,552	0,561
diffAverage_GLCM	**0,726**	**0,724**	0,698	0,694	0,686	0,689	0,674	0,673	0,684	0,682
diffVariance_GLCM	**0,711**	**0,710**	0,657	0,652	0,685	0,688	0,640	0,638	0,673	0,678
diffEntropy_GLCM	**0,723**	**0,721**	0,688	0,678	0,657	0,659	0,664	0,669	0,678	0,681
sumAverage_GLCM	0,649	0,647	0,659	0,658	0,625	0,625	0,624	0,624	0,572	0,572
sumVariance_GLCM	**0,716**	**0,717**	0,678	0,680	0,682	0,682	**0,730**	**0,728**	0,666	0,669
sumEntropy_GLCM	**0,711**	**0,705**	0,624	0,541	0,594	0,482	0,676	0,644	0,628	0,574
angularSecondMoment_GLCM	**0,713**	0,656	0,594	0,541	0,540	0,602	0,619	0,534	0,561	0,537
contrast_GLCM	**0,726**	**0,726**	0,685	0,684	0,683	0,681	0,664	0,672	0,684	0,685
dissimilarity_GLCM	**0,726**	**0,724**	0,698	0,694	0,686	0,689	0,674	0,673	0,684	0,682
InverseDifference_GLCM	**0,716**	**0,716**	0,695	0,688	**0,700**	**0,714**	0,664	0,664	0,660	0,659
NormalisedInverseDifference_GLCM	**0,728**	**0,726**	0,700	0,700	0,686	0,693	0,674	0,676	0,682	0,682
InverseDifferenceMoment_GLCM	**0,714**	**0,717**	0,688	0,680	**0,701**	**0,708**	0,665	0,662	0,668	0,658
NormalisedInverseDifferenceMoment_GLCM	**0,726**	**0,726**	0,685	0,682	0,682	0,677	0,668	0,674	0,687	0,685
inverseVariance_GLCM	**0,730**	**0,715**	0,682	0,690	0,669	0,684	0,651	0,658	0,681	0,687
correlation_GLCM	0,627	0,628	0,582	0,583	0,585	0,588	0,568	0,569	0,605	0,606
Autocorrelation_GLCM	0,662	0,662	0,665	0,666	0,631	0,629	0,638	0,638	0,589	0,590
clustertendency_GLCM	**0,716**	**0,717**	0,678	0,680	0,682	0,682	**0,730**	**0,728**	0,666	0,669
clustershad_GLCMe	0,578	0,579	0,546	0,547	0,493	0,494	0,535	0,535	0,491	0,490
clusterprominence_GLCM	0,647	0,651	0,647	0,645	0,676	0,679	**0,727**	**0,729**	0,682	0,688
infCorr1_GLCM	0,650	0,563	0,610	**0,706**	0,665	**0,712**	0,547	0,592	0,513	0,643
infCorr2_GLCM	0,620	0,544	0,635	**0,726**	0,676	**0,709**	0,488	0,623	0,527	0,653
SRE_GLRLM	**0,724**	**0,729**	**0,706**	0,698	0,677	0,673	0,655	0,673	0,668	0,649
LRE_GLRLM	**0,734**	**0,739**	**0,705**	0,683	0,676	0,654	0,685	0,679	0,671	0,641
LGRE_GLRLM	0,566	0,612	0,521	0,612	0,526	0,600	0,552	0,594	0,545	0,576
HGRE_GLRLM	0,681	0,681	0,672	0,673	0,656	0,656	0,637	0,644	0,585	0,590
SRLGE_GLRLM	0,582	0,617	0,527	0,606	0,529	0,610	0,564	0,603	0,554	0,578
SRHGE_GLRLM	0,685	0,682	0,678	0,676	0,662	0,657	0,646	0,647	0,588	0,593
LRLGE_GLRLM	0,489	0,429	0,518	0,602	0,505	0,573	0,513	0,575	0,473	0,565
LRHGE_GLRLM	0,623	0,656	0,633	0,655	0,627	0,641	0,568	0,599	0,548	0,582
GLNU_GLRLM	**0,758**	**0,758**	**0,729**	**0,727**	**0,712**	**0,708**	**0,714**	**0,711**	**0,709**	**0,708**
GLNU_norm_GLRLM	**0,717**	0,672	0,618	0,487	0,547	0,526	0,660	0,602	0,606	0,544
RLNU_GLRLM	**0,734**	**0,741**	0,695	0,698	**0,704**	0,695	0,694	0,689	0,689	0,694
RLNU_norm_GLRLM	**0,725**	**0,729**	**0,708**	0,696	0,681	0,680	0,655	0,671	0,666	0,649
RP_GLRLM	**0,737**	**0,732**	**0,708**	0,677	0,680	0,656	0,670	0,676	0,668	0,649
GreylevelVariance_GLRLM	**0,736**	**0,738**	**0,741**	**0,738**	0,693	0,698	**0,729**	**0,729**	0,682	0,684
RunlengthVariance_GLRLM	**0,746**	**0,748**	**0,713**	**0,700**	0,676	0,647	0,700	0,674	0,669	0,641
RunEntropy_GLRLM	0,496	0,444	0,538	0,566	0,571	0,581	0,553	0,522	0,520	0,529

Table 2.8: AUC values of features derived from T2W images. In the rows the features and in the columns the datasets. Values greater than 0.7 are highlighted in bold.

significant way.

Correlation feature-output

The results of the correlation coefficient calculated between each feature and the outcome are shown in table 2.11 and 2.12 for the datasets extracted from ADC maps and T2W images respectively. These results show that none of the variables in the datasets are strongly correlated with the output. Indeed, the correlation maxima are around ± 0.4.

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	**0,007**	**0,007**	0,071	0,071	0,069	0,069	0,090	0,090	0,090	0,090
mean_ROI_STAT	0,516	0,516	0,819	0,819	0,777	0,777	0,805	0,805	0,777	0,777
minimum_ROI_STAT	0,375	0,375	0,552	0,552	0,546	0,546	0,583	0,583	0,577	0,577
maximum_ROI_STAT	0,708	0,708	0,415	0,415	0,385	0,385	0,464	0,464	0,437	0,437
Range_STAT	**0,008**	**0,008**	**0,011**	**0,011**	**0,012**	**0,012**	**0,030**	**0,030**	**0,034**	**0,034**
1stPercentile_STAT	0,365	0,365	0,602	0,602	0,595	0,595	0,589	0,589	0,602	0,602
10thPercentile_STAT	0,284	0,284	0,628	0,628	0,564	0,564	0,654	0,654	0,641	0,641
25thPercentile_STAT	0,360	0,360	0,674	0,674	0,674	0,674	0,667	0,667	0,647	0,647
50thPercentile_STAT	0,365	0,365	0,680	0,680	0,694	0,694	0,714	0,714	0,707	0,707
75thPercentile_STAT	0,583	0,583	0,891	0,891	0,840	0,840	0,905	0,905	0,819	0,819
90thPercentile_STAT	0,777	0,777	1,000	1,000	0,993	0,993	0,964	0,964	0,985	0,985
95thPercentile_STAT	0,798	0,798	0,522	0,522	0,415	0,415	0,528	0,528	0,481	0,481
IQR_STAT	0,051	0,051	0,151	0,151	0,148	0,148	0,175	0,175	0,264	0,264
skewness_STAT	**0,017**	**0,017**	0,136	0,136	0,184	0,184	0,122	0,122	0,098	0,098
kurtosis_STAT	0,504	0,504	0,161	0,161	0,131	0,131	0,179	0,179	0,074	0,074
IntensityKurtosis_STAT	0,493	0,493	0,184	0,184	0,141	0,141	0,179	0,179	0,074	0,074
IntensityVariance_STAT	**0,033**	**0,033**	**0,036**	**0,036**	**0,047**	**0,047**	0,090	0,090	0,105	0,105
mean_intensity_IH	0,056	0,056	0,253	0,238	0,276	0,293	0,245	0,238	0,156	0,156
JointMax_GLCM	0,253	0,848	**0,026**	0,905	0,260	0,504	**0,035**	0,641	0,105	0,641
JointAverage_GLCM	0,084	0,084	0,238	0,238	0,310	0,310	0,197	0,191	0,191	0,203
JointVariance_GLCM	0,763	0,763	0,210	0,210	0,084	0,084	0,447	0,447	0,063	0,066
JointEntropy_GLCM	**0,045**	0,805	0,492	0,641	0,577	0,415	0,053	0,920	0,058	0,833
diffAverage_GLCM	**0,024**	**0,024**	0,053	0,056	**0,041**	**0,045**	0,131	0,113	0,087	0,090
diffVariance_GLCM	0,056	0,054	0,161	0,146	0,167	0,179	0,365	0,365	0,210	0,217
diffEntropy_GLCM	**0,032**	**0,032**	0,074	0,074	0,077	0,058	0,173	0,141	0,161	0,122
sumAverage_GLCM	0,084	0,084	0,238	0,238	0,310	0,310	0,197	0,191	0,191	0,203
sumVariance_GLCM	0,819	0,848	0,437	0,447	0,217	0,245	0,528	0,492	0,118	0,126
sumEntropy_GLCM	0,615	0,777	0,293	0,667	0,231	0,905	0,210	0,447	0,179	0,365
angularSecondMoment_GLCM	**0,047**	0,540	0,053	0,978	0,337	0,654	**0,007**	0,224	**0,017**	0,355
contrast_GLCM	**0,035**	**0,033**	0,080	0,087	0,087	0,087	0,167	0,167	0,146	0,141
dissimilarity_GLCM	**0,024**	**0,024**	0,053	0,056	**0,041**	**0,045**	0,131	0,113	0,087	0,090
InverseDifference_GLCM	**0,018**	**0,016**	**0,029**	**0,019**	**0,013**	**0,016**	**0,036**	**0,023**	**0,022**	**0,023**
NormalisedInverseDifference_GLCM	**0,023**	**0,023**	0,051	**0,049**	**0,030**	**0,039**	0,109	0,113	0,071	0,074
InverseDifferenceMoment_GLCM	**0,018**	**0,013**	**0,019**	**0,011**	**0,018**	**0,009**	**0,036**	**0,012**	**0,020**	**0,013**
NormalisedInverseDifferenceMoment_GLCM	**0,032**	**0,032**	0,077	0,087	0,084	0,087	0,167	0,167	0,122	0,131
inverseVariance_GLCM	**0,026**	**0,016**	**0,024**	**0,008**	**0,045**	**0,006**	0,231	**0,036**	0,080	**0,009**
correlation_GLCM	**0,038**	**0,040**	0,276	0,276	0,301	0,301	0,405	0,405	0,395	0,395
Autocorrelation_GLCM	0,146	0,146	0,224	0,224	0,268	0,268	0,167	0,167	0,126	0,126
clustertendency_GLCM	0,819	0,848	0,437	0,447	0,217	0,245	0,528	0,492	0,118	0,126
clustershad_GLCMe	**0,007**	**0,007**	0,365	0,375	0,318	0,337	0,224	0,238	0,276	0,293
clusterprominence_GLCM	0,602	0,602	0,552	0,528	0,346	0,337	0,978	0,934	0,492	0,492
infCorr1_GLCM	**0,045**	0,328	0,891	0,141	0,540	0,087	0,437	0,680	0,667	0,365
infCorr2_GLCM	**0,049**	0,310	0,577	0,146	0,355	0,061	0,721	0,516	0,791	0,346
SRE_GLRLM	**0,013**	**0,012**	0,056	0,101	**0,005**	**0,035**	**0,036**	**0,035**	**0,016**	0,056
LRE_GLRLM	**0,006**	**0,007**	**0,014**	**0,041**	**0,004**	**0,025**	**0,010**	**0,015**	**0,005**	**0,015**
LGRE_GLRLM	0,654	0,862	0,680	0,301	0,293	0,094	0,721	0,395	0,615	0,385
HGRE_GLRLM	0,074	0,071	0,136	0,131	0,126	0,126	0,173	0,161	0,122	0,118
SRLGE_GLRLM	0,763	0,993	0,641	0,253	0,245	0,069	0,654	0,437	0,458	0,301
SRHGE_GLRLM	0,066	0,054	0,126	0,126	0,122	0,156	0,122	0,118	0,090	0,094
LRLGE_GLRLM	0,355	0,694	0,791	0,327	0,540	0,173	0,819	0,615	0,949	0,707
LRHGE_GLRLM	0,293	0,146	0,516	0,268	0,641	0,301	0,891	0,385	0,862	0,318
GLNU_GLRLM	**0,006**	**0,007**	**0,049**	**0,045**	**0,035**	**0,036**	0,061	0,053	0,058	0,051
GLNU_norm_GLRLM	0,934	0,405	0,934	0,993	0,365	0,905	0,318	0,654	0,131	0,694
RLNU_GLRLM	**0,007**	**0,006**	0,098	0,069	0,094	0,069	0,090	0,090	0,090	0,080
RLNU_norm_GLRLM	**0,014**	**0,013**	0,058	0,122	**0,004**	**0,031**	**0,038**	0,056	**0,020**	0,051
RP_GLRLM	**0,009**	**0,010**	**0,019**	0,071	**0,005**	**0,033**	**0,014**	**0,018**	**0,006**	**0,031**
GreylevelVariance_GLRLM	0,891	0,934	0,447	0,426	0,173	0,098	0,654	0,492	0,105	0,080
RunlengthVariance_GLRLM	**0,004**	**0,006**	**0,007**	**0,033**	**0,005**	**0,025**	**0,005**	**0,010**	**0,006**	**0,009**
RunEntropy_GLRLM	0,061	0,056	0,301	0,426	0,276	0,284	0,151	0,224	0,101	0,447

Table 2.9: Results of the Mann–Whitney U test of ADC map features in terms of p-value, for each of the 10 datasets. Values lower than 0.05 are highlighted in bold.

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	**0,008**	**0,008**	0,074	0,074	0,069	0,069	0,087	0,087	0,090	0,090
JointMax_GLCM	0,094	0,319	0,805	0,964	0,504	0,437	0,577	0,862	0,721	0,707
JointAverage_GLCM	**0,020**	**0,022**	0,058	0,058	0,203	0,210	0,224	0,217	0,694	0,694
JointVariance_GLCM	**0,003**	**0,003**	**0,020**	**0,021**	0,080	0,077	**0,015**	**0,016**	0,146	0,146
JointEntropy_GLCM	**0,009**	0,098	0,492	0,654	0,949	0,481	0,253	0,848	0,694	0,628
diffAverage_GLCM	**0,003**	**0,004**	**0,031**	**0,035**	0,066	0,061	0,066	0,066	**0,049**	0,056
diffVariance_GLCM	**0,007**	**0,007**	0,118	0,122	**0,043**	**0,039**	0,113	0,122	**0,043**	**0,036**
diffEntropy_GLCM	**0,004**	**0,004**	**0,045**	**0,049**	0,156	0,109	0,080	0,069	0,058	**0,049**
sumAverage_GLCM	**0,020**	**0,022**	0,058	0,058	0,203	0,210	0,224	0,217	0,694	0,694
sumVariance_GLCM	**0,009**	**0,009**	**0,031**	**0,031**	0,090	0,098	**0,013**	**0,014**	0,203	0,191
sumEntropy_GLCM	**0,007**	**0,009**	0,141	0,415	0,385	0,833	0,061	0,109	0,437	0,470
angularSecondMoment_GLCM	**0,009**	0,069	0,284	1,000	0,833	0,470	0,301	0,805	0,654	0,833
contrast_GLCM	**0,004**	**0,004**	0,061	0,058	0,063	0,066	0,087	0,063	**0,049**	**0,049**
dissimilarity_GLCM	**0,003**	**0,004**	**0,031**	**0,035**	0,066	0,061	0,066	0,066	**0,049**	0,056
InverseDifference_GLCM	**0,006**	**0,006**	**0,031**	**0,038**	**0,043**	**0,026**	0,098	0,101	0,122	0,136
NormalisedInverseDifference_GLCM	**0,003**	**0,003**	**0,029**	**0,026**	0,066	0,056	0,066	0,066	0,056	0,053
InverseDifferenceMoment_GLCM	**0,007**	**0,005**	**0,045**	**0,049**	**0,041**	**0,027**	0,098	0,098	0,098	0,113
NormalisedInverseDifferenceMoment_GLCM	**0,004**	**0,004**	0,056	0,056	0,071	0,084	0,074	0,058	**0,043**	**0,047**
inverseVariance_GLCM	**0,004**	**0,007**	0,051	**0,043**	0,161	0,087	0,191	0,126	0,063	0,051
correlation_GLCM	0,087	0,084	0,415	0,395	0,310	0,276	0,447	0,458	0,184	0,179
Autocorrelation_GLCM	**0,015**	**0,015**	**0,049**	0,051	0,197	0,203	0,156	0,156	0,540	0,540
clustertendency_GLCM	**0,009**	**0,009**	**0,031**	**0,031**	0,090	0,098	**0,013**	**0,014**	0,203	0,191
clustershad_GLCMe	0,173	0,185	0,680	0,654	0,735	0,763	0,805	0,819	0,694	0,694
clusterprominence_GLCM	0,054	**0,045**	**0,038**	**0,043**	0,077	0,069	**0,013**	**0,012**	0,087	0,074
infCorr1_GLCM	0,063	0,328	0,260	0,087	0,197	0,087	0,470	0,395	0,876	0,268
infCorr2_GLCM	0,113	0,426	0,151	**0,027**	0,131	0,058	0,735	0,238	0,934	0,224
SRE_GLRLM	**0,004**	**0,004**	**0,035**	**0,025**	**0,033**	**0,018**	0,161	0,118	0,173	0,301
LRE_GLRLM	**0,002**	**0,002**	**0,025**	**0,049**	**0,033**	**0,030**	0,063	0,101	0,118	0,301
LGRE_GLRLM	0,735	0,708	0,993	0,318	1,000	0,564	0,641	0,346	0,763	0,437
HGRE_GLRLM	**0,008**	**0,008**	**0,041**	**0,043**	0,087	0,105	0,151	0,131	0,589	0,540
SRLGE_GLRLM	0,934	0,667	0,920	0,405	0,978	0,458	0,667	0,293	0,735	0,405
SRHGE_GLRLM	**0,006**	**0,007**	**0,039**	**0,041**	0,084	0,098	0,131	0,118	0,589	0,552
LRLGE_GLRLM	0,293	0,920	0,833	0,346	0,707	0,833	0,833	0,395	0,777	0,458
LRHGE_GLRLM	**0,049**	**0,022**	0,113	0,063	0,231	0,173	0,504	0,385	0,763	0,458
GLNU_GLRLM	**0,005**	**0,005**	0,058	0,056	0,071	0,074	0,074	0,080	0,071	0,069
GLNU_norm_GLRLM	**0,009**	**0,038**	0,173	0,589	0,694	0,920	0,118	0,293	0,447	0,694
RLNU_GLRLM	**0,011**	**0,007**	0,084	0,074	0,071	0,084	0,098	0,105	0,101	0,080
RLNU_norm_GLRLM	**0,004**	**0,004**	**0,038**	**0,024**	**0,030**	**0,015**	0,161	0,122	0,184	0,301
RP_GLRLM	**0,002**	**0,003**	**0,022**	0,061	**0,029**	**0,033**	0,101	0,109	0,151	0,260
GreylevelVariance_GLRLM	**0,005**	**0,005**	**0,008**	**0,009**	0,069	0,063	**0,024**	**0,022**	0,109	0,094
RunlengthVariance_GLRLM	**0,001**	**0,002**	**0,023**	**0,025**	**0,033**	**0,036**	**0,038**	0,113	0,113	0,260
RunEntropy_GLRLM	0,964	0,540	0,949	0,805	0,437	0,602	0,819	0,964	0,876	0,876

Table 2.10: Results of the Mann–Whitney U test of T2w image features in terms of p-value, for each of the 10 datasets. Values lower than 0.05 are highlighted in bold.

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	0,383	0,383	0,337	0,337	0,337	0,337	0,334	0,334	0,334	0,334
mean_ROI_STAT	-0,112	-0,112	-0,078	-0,078	-0,074	-0,074	-0,077	-0,077	-0,074	-0,074
minimum_ROI_STAT	-0,199	-0,199	-0,188	-0,188	-0,202	-0,202	-0,186	-0,186	-0,194	-0,194
maximum_ROI_STAT	0,062	0,062	0,099	0,099	0,101	0,101	0,089	0,089	0,097	0,097
Range_STAT	0,357	0,357	0,342	0,342	0,346	0,346	0,315	0,315	0,315	0,315
1stPercentile_STAT	-0,189	-0,189	-0,170	-0,170	-0,184	-0,184	-0,170	-0,170	-0,178	-0,178
10thPercentile_STAT	-0,169	-0,169	-0,145	-0,145	-0,147	-0,147	-0,142	-0,142	-0,141	-0,141
25thPercentile_STAT	-0,154	-0,154	-0,111	-0,111	-0,108	-0,108	-0,111	-0,111	-0,110	-0,110
50thPercentile_STAT	-0,133	-0,133	-0,092	-0,092	-0,088	-0,088	-0,092	-0,092	-0,087	-0,087
75thPercentile_STAT	-0,080	-0,080	-0,050	-0,050	-0,043	-0,043	-0,049	-0,049	-0,047	-0,047
90thPercentile_STAT	-0,021	-0,021	-0,010	-0,010	-0,008	-0,008	0,003	0,003	0,005	0,005
95thPercentile_STAT	0,043	0,043	0,068	0,068	0,074	0,074	0,065	0,065	0,075	0,075
IQR_STAT	0,269	0,269	0,191	0,191	0,197	0,197	0,193	0,193	0,185	0,185
skewness_STAT	0,265	0,265	0,075	0,075	0,058	0,058	0,161	0,161	0,156	0,156
kurtosis_STAT	0,091	0,091	0,064	0,064	0,039	0,039	0,164	0,164	0,178	0,178
IntensityKurtosis_STAT	0,088	0,088	0,056	0,056	0,030	0,030	0,161	0,161	0,174	0,174
IntensityVariance_STAT	0,290	0,290	0,246	0,246	0,245	0,245	0,241	0,241	0,235	0,235
mean_intensity_IH	-0,213	-0,214	-0,065	-0,065	-0,039	-0,037	-0,081	-0,081	-0,097	-0,096
JointMax_GLCM	0,185	0,013	0,101	-0,004	0,057	-0,213	0,175	-0,041	0,160	-0,068
JointAverage_GLCM	-0,196	-0,196	-0,047	-0,047	-0,034	-0,033	-0,071	-0,071	-0,092	-0,091
JointVariance_GLCM	-0,118	-0,120	-0,172	-0,172	-0,185	-0,183	-0,215	-0,217	-0,237	-0,236
JointEntropy_GLCM	-0,259	-0,046	0,077	0,242	0,119	0,267	-0,186	0,128	-0,179	0,149
diffAverage_GLCM	-0,328	-0,329	-0,288	-0,286	-0,302	-0,303	-0,306	-0,307	-0,326	-0,326
diffVariance_GLCM	-0,271	-0,271	-0,219	-0,218	-0,234	-0,232	-0,244	-0,245	-0,264	-0,265
diffEntropy_GLCM	-0,323	-0,325	-0,245	-0,244	-0,248	-0,246	-0,254	-0,264	-0,273	-0,281
sumAverage_GLCM	-0,196	-0,196	-0,047	-0,047	-0,034	-0,033	-0,071	-0,071	-0,092	-0,091
sumVariance_GLCM	-0,034	-0,037	-0,133	-0,134	-0,146	-0,145	-0,192	-0,194	-0,215	-0,214
sumEntropy_GLCM	-0,091	-0,069	0,023	0,117	0,048	0,140	-0,115	-0,012	-0,159	-0,062
angularSecondMoment_GLCM	0,270	0,138	0,016	-0,157	-0,020	-0,227	0,239	0,048	0,232	-0,014
contrast_GLCM	-0,289	-0,289	-0,255	-0,254	-0,281	-0,280	-0,284	-0,284	-0,311	-0,312
dissimilarity_GLCM	-0,328	-0,329	-0,288	-0,286	-0,302	-0,303	-0,306	-0,307	-0,326	-0,326
InverseDifference_GLCM	0,344	0,343	0,333	0,329	0,343	0,357	0,331	0,335	0,343	0,340
NormalisedInverseDifference_GLCM	0,336	0,336	0,295	0,293	0,307	0,308	0,310	0,311	0,328	0,328
InverseDifferenceMoment_GLCM	0,342	0,339	0,335	0,338	0,340	0,372	0,326	0,336	0,340	0,343
NormalisedInverseDifferenceMoment_GLCM	0,300	0,300	0,259	0,258	0,283	0,282	0,285	0,285	0,312	0,313
inverseVariance_GLCM	0,338	0,343	0,279	0,369	0,259	0,357	0,248	0,321	0,287	0,357
correlation_GLCM	0,308	0,308	0,251	0,251	0,249	0,249	0,238	0,240	0,230	0,232
Autocorrelation_GLCM	-0,200	-0,201	-0,071	-0,072	-0,062	-0,062	-0,097	-0,098	-0,115	-0,115
clustertendency_GLCM	-0,034	-0,037	-0,133	-0,134	-0,146	-0,145	-0,192	-0,194	-0,215	-0,214
clustershad_GLCMe	0,285	0,287	0,083	0,083	0,091	0,088	0,108	0,106	0,071	0,069
clusterprominence_GLCM	0,030	0,028	-0,137	-0,139	-0,161	-0,160	-0,138	-0,143	-0,186	-0,186
infCorr1_GLCM	-0,268	-0,132	0,056	0,241	0,146	0,312	-0,125	0,119	-0,085	0,186
infCorr2_GLCM	0,285	0,150	0,040	-0,125	-0,049	-0,253	0,144	-0,044	0,101	-0,140
SRE_GLRLM	-0,356	-0,352	-0,335	-0,269	-0,401	-0,377	-0,298	-0,310	-0,329	-0,267
LRE_GLRLM	0,381	0,374	0,386	0,314	0,409	0,392	0,335	0,339	0,351	0,302
LGRE_GLRLM	-0,080	-0,130	-0,198	-0,242	-0,263	-0,338	-0,178	-0,230	-0,197	-0,239
HGRE_GLRLM	-0,223	-0,231	-0,104	-0,110	-0,089	-0,091	-0,099	-0,114	-0,126	-0,129
SRLGE_GLRLM	-0,110	-0,141	-0,218	-0,245	-0,289	-0,358	-0,223	-0,235	-0,230	-0,259
SRHGE_GLRLM	-0,237	-0,238	-0,122	-0,118	-0,113	-0,100	-0,125	-0,129	-0,160	-0,140
LRLGE_GLRLM	0,045	-0,080	-0,110	-0,202	-0,148	-0,264	-0,053	-0,164	-0,080	-0,159
LRHGE_GLRLM	-0,132	-0,191	-0,011	-0,069	0,009	-0,046	0,050	-0,024	0,045	-0,058
GLNU_GLRLM	0,380	0,381	0,363	0,374	0,367	0,375	0,340	0,351	0,347	0,358
GLNU_norm_GLRLM	0,066	-0,047	-0,040	-0,176	0,004	-0,165	0,108	-0,023	0,158	-0,002
RLNU_GLRLM	0,386	0,384	0,311	0,326	0,310	0,324	0,315	0,324	0,311	0,324
RLNU_norm_GLRLM	-0,355	-0,350	-0,329	-0,262	-0,400	-0,374	-0,298	-0,306	-0,330	-0,264
RP_GLRLM	-0,372	-0,364	-0,367	-0,297	-0,406	-0,383	-0,342	-0,337	-0,358	-0,292
GreylevelVariance_GLRLM	-0,109	-0,117	-0,236	-0,242	-0,256	-0,263	-0,204	-0,219	-0,241	-0,249
RunlengthVariance_GLRLM	0,397	0,388	0,397	0,325	0,407	0,396	0,344	0,341	0,358	0,314
RunEntropy_GLRLM	0,302	0,267	0,274	0,285	0,311	0,298	0,277	0,273	0,290	0,245

Table 2.11: Results of the correlation between each feature and the outcome, for each of the 10 datasets coming from ADC maps.

	3D - GAUSSIAN - 0,5		2D - GAUSSIAN - 0,5		2D - NOBLUR - 0,5		2D - GAUSSIAN - 0,31		2D - NOBLUR - 0,31	
	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin	32bin	64bin
ROI_volume(mm3)	0,383	0,383	0,336	0,336	0,336	0,336	0,333	0,333	0,333	0,333
JointMax_GLCM	0,072	0,071	0,131	-0,017	-0,087	-0,196	0,118	0,040	0,085	-0,113
JointAverage_GLCM	-0,207	-0,207	-0,235	-0,235	-0,184	-0,185	-0,186	-0,187	-0,127	-0,127
JointVariance_GLCM	-0,359	-0,359	-0,326	-0,328	-0,315	-0,316	-0,336	-0,337	-0,293	-0,292
JointEntropy_GLCM	-0,312	-0,178	-0,071	0,189	0,108	0,251	-0,209	-0,021	-0,126	0,120
diffAverage_GLCM	-0,380	-0,380	-0,351	-0,349	-0,327	-0,330	-0,315	-0,317	-0,315	-0,314
diffVariance_GLCM	-0,360	-0,358	-0,293	-0,294	-0,291	-0,292	-0,247	-0,247	-0,292	-0,291
diffEntropy_GLCM	-0,382	-0,381	-0,337	-0,335	-0,316	-0,310	-0,299	-0,301	-0,324	-0,322
sumAverage_GLCM	-0,207	-0,207	-0,235	-0,235	-0,184	-0,185	-0,186	-0,187	-0,127	-0,127
sumVariance_GLCM	-0,314	-0,314	-0,290	-0,292	-0,269	-0,269	-0,323	-0,324	-0,269	-0,268
sumEntropy_GLCM	-0,240	-0,224	-0,178	-0,042	-0,128	0,014	-0,206	-0,169	-0,169	-0,096
angularSecondMoment_GLCM	0,186	0,131	0,150	-0,134	-0,013	-0,241	0,154	0,073	0,181	-0,037
contrast_GLCM	-0,370	-0,369	-0,336	-0,334	-0,298	-0,299	-0,301	-0,302	-0,297	-0,296
dissimilarity_GLCM	-0,380	-0,380	-0,351	-0,349	-0,327	-0,330	-0,315	-0,317	-0,315	-0,314
InverseDifference_GLCM	0,371	0,367	0,362	0,353	0,353	0,371	0,315	0,321	0,331	0,326
NormalisedInverseDifference_GLCM	0,380	0,380	0,353	0,350	0,334	0,337	0,316	0,318	0,318	0,317
InverseDifferenceMoment_GLCM	0,369	0,358	0,359	0,350	0,355	0,372	0,313	0,317	0,333	0,328
NormalisedInverseDifferenceMoment_GLCM	0,371	0,371	0,337	0,336	0,305	0,306	0,303	0,304	0,299	0,298
inverseVariance_GLCM	0,393	0,367	0,320	0,348	0,336	0,338	0,294	0,301	0,335	0,353
correlation_GLCM	0,258	0,259	0,180	0,182	0,176	0,179	0,140	0,144	0,191	0,193
Autocorrelation_GLCM	-0,213	-0,214	-0,231	-0,230	-0,185	-0,186	-0,208	-0,209	-0,139	-0,138
clustertendency_GLCM	-0,314	-0,314	-0,290	-0,292	-0,269	-0,269	-0,323	-0,324	-0,269	-0,268
clustershad_GLCMe	0,081	0,080	0,005	0,010	-0,057	-0,062	0,027	0,028	-0,050	-0,052
clusterprominence_GLCM	-0,252	-0,251	-0,243	-0,241	-0,252	-0,252	-0,327	-0,327	-0,256	-0,256
infCorr1_GLCM	-0,250	-0,084	0,215	0,379	0,295	0,340	-0,064	0,185	0,014	0,265
infCorr2_GLCM	0,213	0,051	-0,217	-0,379	-0,315	-0,359	-0,012	-0,178	-0,049	-0,251
SRE_GLRLM	-0,378	-0,382	-0,366	-0,339	-0,309	-0,327	-0,296	-0,307	-0,329	-0,307
LRE_GLRLM	0,390	0,392	0,361	0,347	0,303	0,298	0,323	0,326	0,331	0,299
LGRE_GLRLM	-0,047	-0,113	-0,025	-0,214	-0,056	-0,182	-0,054	-0,154	-0,009	-0,076
HGRE_GLRLM	-0,238	-0,242	-0,253	-0,251	-0,205	-0,208	-0,219	-0,221	-0,151	-0,154
SRLGE_GLRLM	-0,106	-0,148	-0,042	-0,223	-0,083	-0,197	-0,099	-0,166	-0,033	-0,089
SRHGE_GLRLM	-0,252	-0,250	-0,271	-0,260	-0,210	-0,210	-0,239	-0,232	-0,161	-0,160
LRLGE_GLRLM	0,106	0,001	0,048	-0,156	0,021	-0,120	0,053	-0,049	0,065	-0,014
LRHGE_GLRLM	-0,154	-0,203	-0,172	-0,214	-0,180	-0,203	-0,089	-0,159	-0,085	-0,128
GLNU_GLRLM	0,427	0,432	0,390	0,397	0,376	0,377	0,367	0,378	0,372	0,381
GLNU_norm_GLRLM	0,192	0,120	0,161	-0,033	0,058	-0,145	0,160	0,089	0,154	0,047
RLNU_GLRLM	0,366	0,374	0,307	0,322	0,312	0,323	0,297	0,313	0,307	0,320
RLNU_norm_GLRLM	-0,377	-0,382	-0,367	-0,339	-0,311	-0,330	-0,296	-0,307	-0,324	-0,304
RP_GLRLM	-0,390	-0,391	-0,367	-0,341	-0,309	-0,311	-0,320	-0,324	-0,332	-0,303
GreylevelVariance_GLRLM	-0,360	-0,361	-0,364	-0,358	-0,314	-0,318	-0,361	-0,359	-0,298	-0,303
RunlengthVariance_GLRLM	0,394	0,395	0,354	0,355	0,300	0,272	0,335	0,332	0,323	0,284
RunEntropy_GLRLM	0,070	0,024	0,120	0,161	0,156	0,233	0,126	0,089	0,106	0,103

Table 2.12: Results of the correlation between each feature and the outcome, for each of the 10 datasets coming from T2W images.

2.5.2 Multiparametric analysis

In this section the results related to the multiparametric analysis will be shown. Note that, unlike the univariate analysis, this time only datasets with voxel spacing 0.5mm and with Gaussian filter (see table 2.2, dataset number 1, 2, 3 and 4) are tested, as their univariate analysis showed better results than the others.

Minimum Redundancy Maximum Relevance

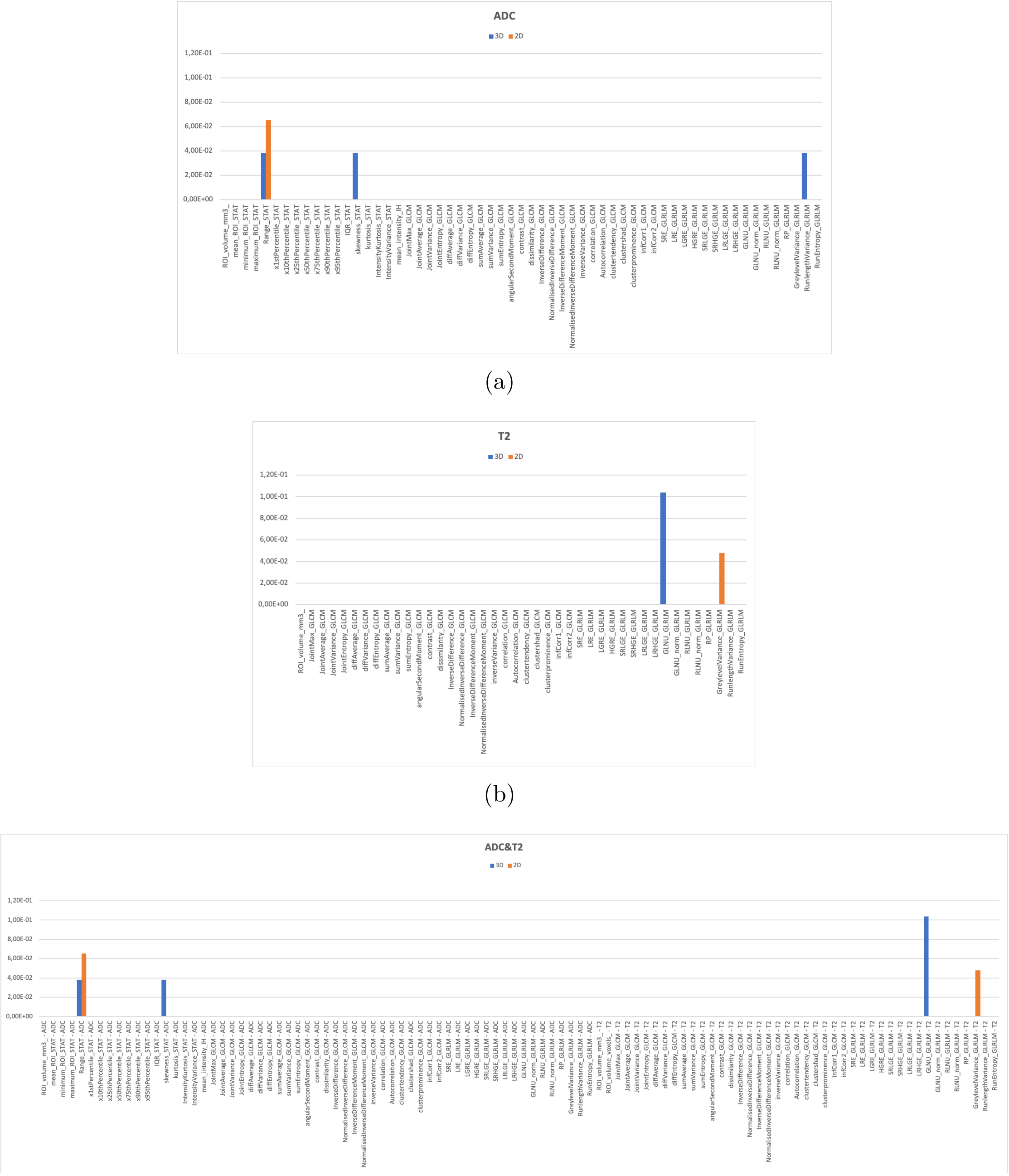
Bar diagrams in figure 2.7 show the score that the MRMR algorithm has attributed to each feature in the three type of dataset, ADC (a), T2 (b) and ADC-T2 joined (c). Features considered most predictive and minimally correlated with each other are evident, unlike the rest of the features that have a value close to zero (less than 10^{-16}). Therefore, table 2.13 shows the selected features for each dataset, written in descending order of score.
What can be seen is that, in all cases, the MRMR algorithm allows to obtain a very low number of features, which underlines their high predictive power. Furthermore, features selected in the case of the ADC-T2 joined dataset, correspond to the features selected by the two datasets, ADC and T2, analyzed separately, with the exception of the 3D case in which one of the features selected by the ADC dataset, "RunlengthVariance_GLRLM", is not present.

	ADC	T2	ADC-T2 joined
3D	skewness_STAT, RunlengthVariance_GLRLM, Range_STAT	GLNU_GLRLM	GLNU_GLRLM-T2, skewness_STAT-ADC, Range_STAT-ADC
2D	range_STAT	GreylevelVariance_GLRLM	Range_STAT-ADC, GreylevelVariance_GLRLM-T2

Table 2.13: Feature selected by MRMR algorithm, in the case of 3D and 2D datasets with 0.5mm voxel spacing and 32bins.

Genetic Algorithm

Figure 2.6 shows the results obtained by the Genetic Algorithm (GA), in the case with svm-dependent fitness formulation with polynomial and Gaussian kernel, in terms of fitness value and number of selected features, for ADC, T2 and ADC-T2 joined datasets. Figure 2.7 shows the corresponing results of the GA with MIQ-based fitness formulation.

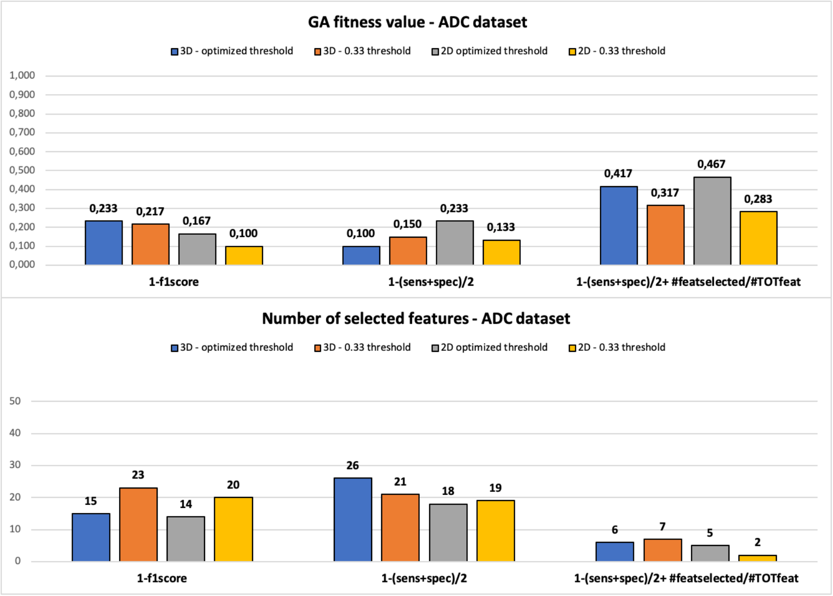


(a)

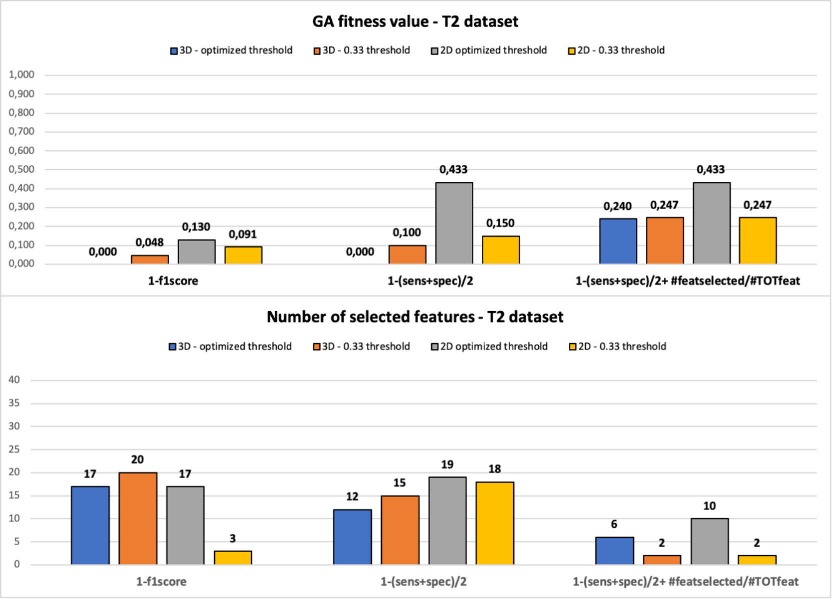
(b)

(c)

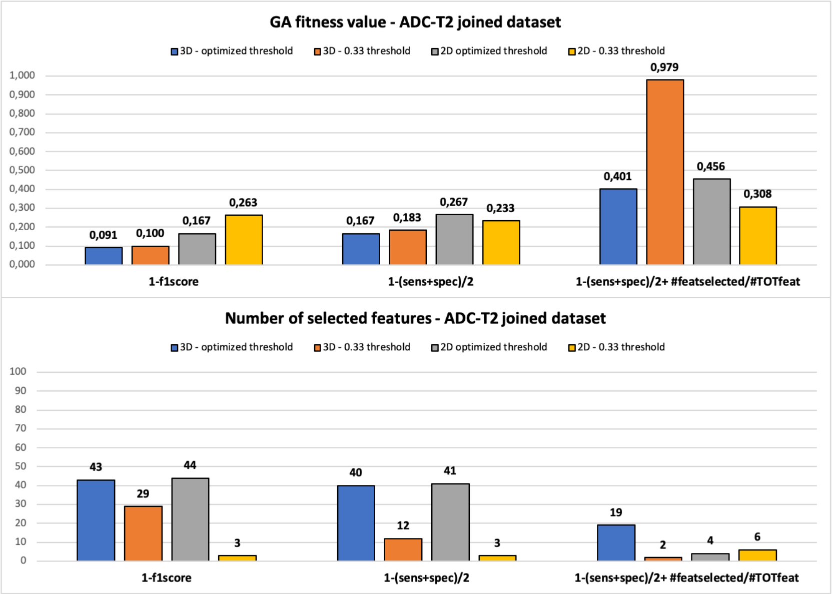
Figure 2.7: MRMR score of 3D (blue) and 2D (orange) features coming from ADC (a), T2 (b) and ADC-T2 joined (c) datasets.



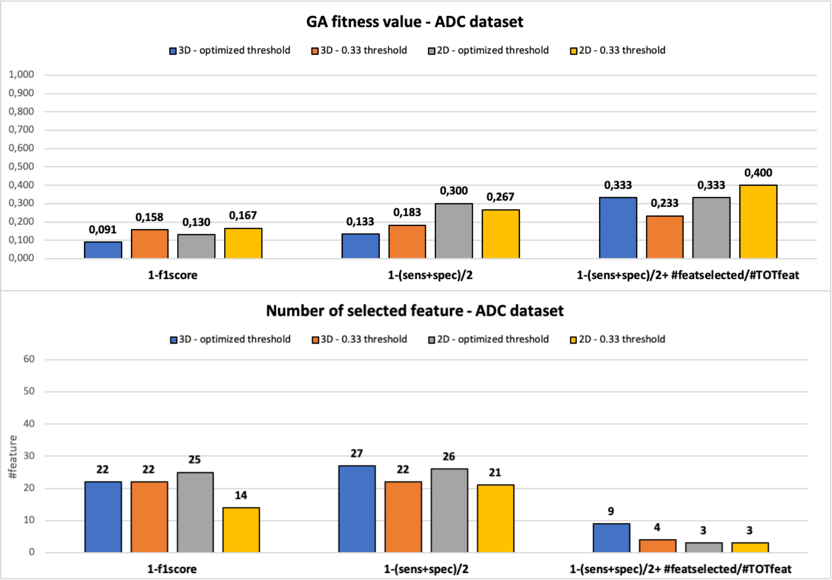
((a)) GA results - fitness formulation with svm and polynomial kernel - ADC dataset.



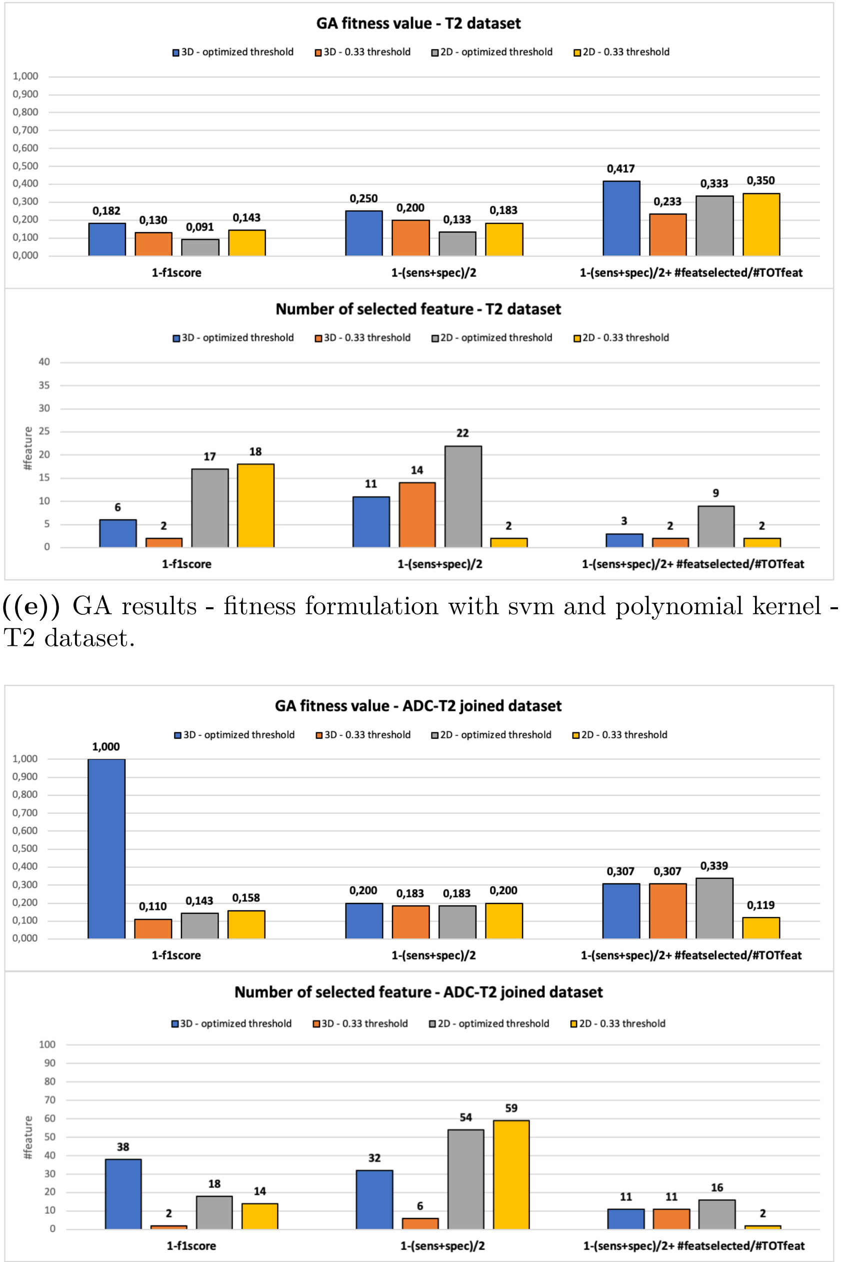
((b)) Ga results for fitness formulation with svm and polynomial kernel - T2 dataset.



((c)) GA results - fitness formulation with svm and polynomial kernel - ADC-T2 joined dataset.



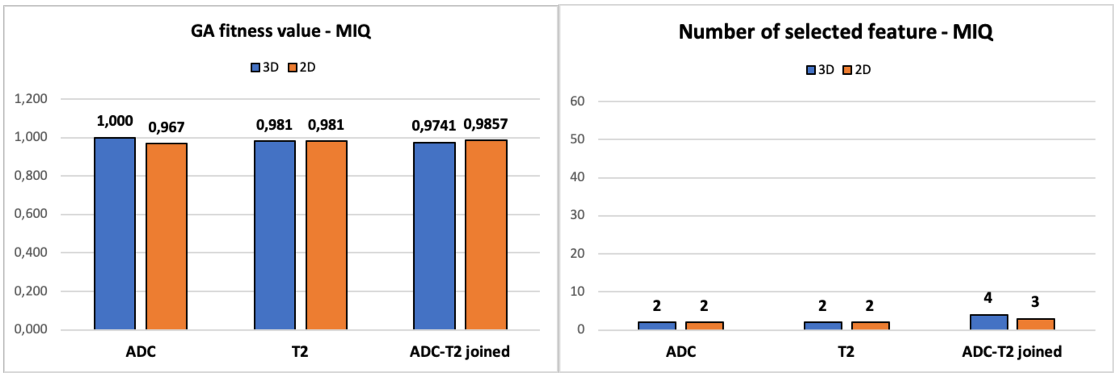
((d)) Ga results - fitness formulation with svm and Gaussian kernel - ADC dataset.



((e)) GA results - fitness formulation with svm and polynomial kernel - T2 dataset.

((f)) GA results - fitness formulation with svm and Gaussian kernel - ADC-T2 joined dataset.

Figure 2.6: GA results in the case of svm-dependent fitness, with polynomial (a,b,c) and Gaussian (d,e,f) kernel, in terms of obtained fitness values and number of selected features, for the ADC, T2 and ADC-T2 joined datasets.



((g)) ((h))

Figure 2.7: GA results in the case with MIQ-based fitness formulation, in terms of obtained fitness values (a) and number of selected features (b), for the ADC, T2 and ADC-T2 joined datasets.

In the ADC dataset, the highest svm-dependent fitness values are those related to the penalty term formulation, that takes into account the percentage of selected features. Therefore, it is important to consider this term that should be subtracted in order to compare this fitness value with the others, obtained from the other two fitness formulation tested.
With regard to the number of features, for fitness $1 - f1score$ and $1 - \frac{sens+spec}{2}$, it is around twenty in the case of ADC and T2 datasets, and increases up to 59 in the case of the ADC-T2 joined dataset. For the fitness with penalty term, the number of features never exceeds ten for ADC and T2 datasets, and slightly increases for the ADC-T2 dataset, up to a maximum of 19 selected features. From this it is clear that the penalty term does not cause a decrease in FS performance and leads to an important decrease in the cardinality of the selected feature subset.
The best results in the svm-dependent fitness case are those of the 3D dataset with optimized classification threshold and fitness without penalty: both $1 - f1score$ and $1 - \frac{sens+spec}{2}$ obtain a null value which corresponds to perfect performances, i.e. f1score = 100% and accuracy = 100%, respectively. The worst svm-dependent case, on the other hand, is that of the 3D ADC-T2 joined dataset with optimized threshold, which obtains a fitness value equal to one, i.e. f1score=0%.
Regarding the formulation that is not dependent on the classifier (figure 2.7), i.e. MIQ-based, the fitness value is particularly high. It should be noted that this formulation subtracts one minus the average MIQ attributed to each of the features selected by the GA. These values, in the best cases, have an order of magnitude that is around 10^{-1} and 10^{-2} (as it can be seen in section 2.5.2). Therefore, even if

in this case the aim is to minimize the fitness value, the achievement of values close to one does not necessarily means bad performances of the feature subset. The best and worst results in terms of fitness value are those of the ADC dataset, 2D and 3D respectively. In all cases, the number of features is particularly low, always less than five.

Affinity Propagation

Table 2.14 contains the 3D and 2D features selected by the AP clustering algorithm for the ADC, T2 and ADC-T2 joined datasets. Table 2.15 shows the final number of features selected and the percentage of selected features out of the total of selectable features (59 for ADC, 42 for T2 and 100 for ADC-T2 joined) for each dataset.
At least 40% of the features are always selected, which results in the number of features always greater than or equal to twenty, a particularly high amount when compared with the results obtained in the previous section with the GA.

	ADC	T2	ADC-T2 joined
3D	24 (40%)	20 (47%)	45 (45%)
2D	29 (49%)	21 (50%)	49 (49%)

Table 2.15: Number of feature selected by AP algorithm (% of features selected).

2.5.3 Classifier construction

1st FS method: AUC ranking (only features with AUC>70%)

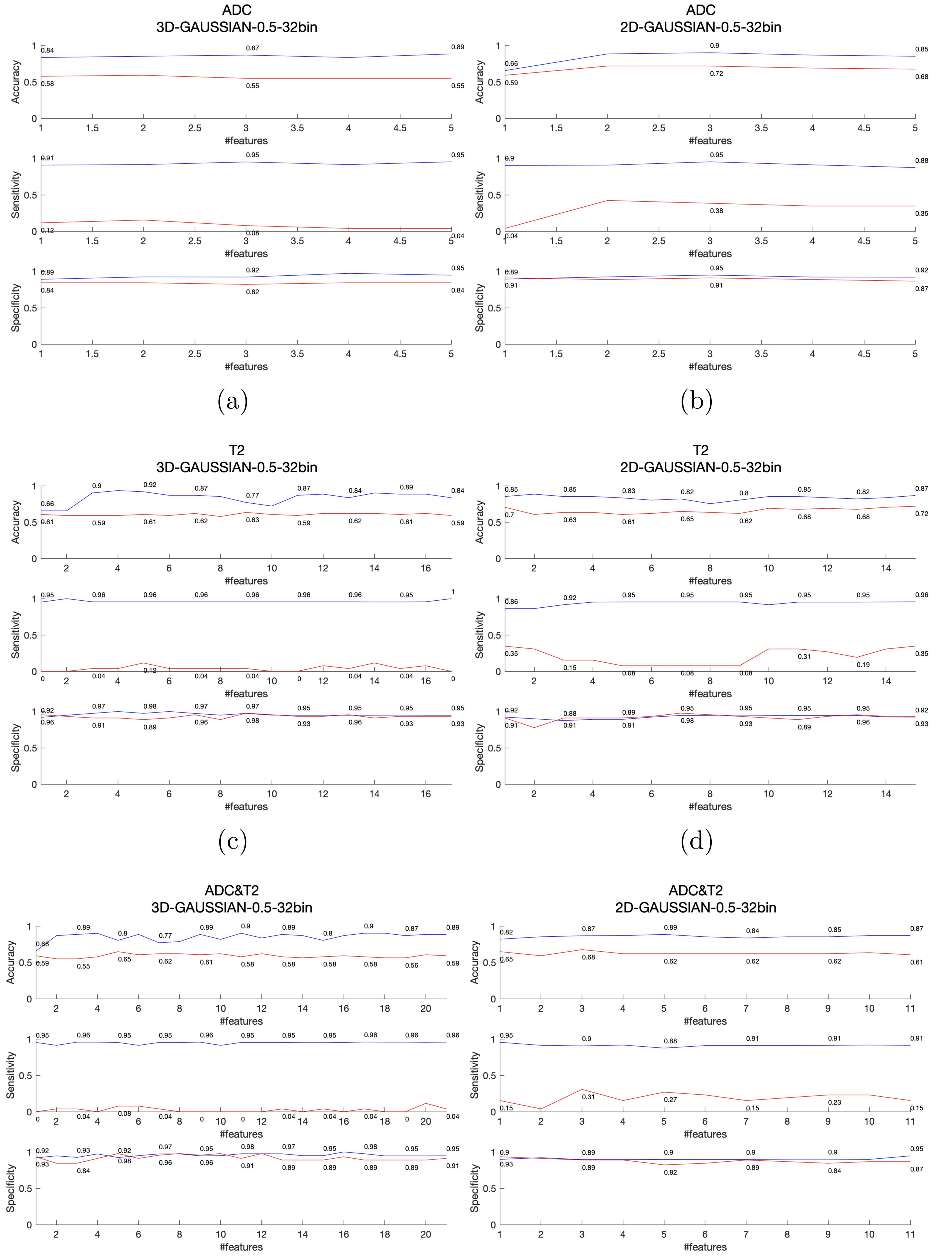
Figure 2.8 shows the performances of the svm classifier with polynomial kernel, in terms of accuracy, sensitivity and specificity. As mentioned in section 2.4.2, 7-fold crossvalidation and an increasing number of features (considering only those with AUC value greater then 0.7) are used, added in descending order of AUC value. Specifically, subfigures (a) and (b) refer to ADC dataset, 3D and 2D respectively; (c) and (d) to T2 dataset, 3D and 2D, respectively; lastly, (e) and (f) to ADC-T2 joined dataset, 3D and 2D respectively. From the obtained performances, it is not possible to identify an overfitting threshold, probably due to an excessively low number of features considered.

2nd FS method: AUC ranking (all features)

Given the unsatisfactory results previously obtained using only the features with AUC>70%, all the features are used. Performances of the svm classifier with polynomial and Gaussian kernel, and of the random forest with 100 trees, are shown in

ADC			T2			ADC-T2 joined		
	3D	2D		3D	2D		3D	2D
ROI_volume_mm3_		X	ROI_volume_mm3_	X	X	ROI_volume_mm3_ - ADC		
mean_ROI_STAT	X	X	JointMax_GLCM	X		mean_ROI_STAT - ADC		X
minimum_ROI_STAT	X		JointAverage_GLCM	X	X	minimum_ROI_STAT - ADC	X	
maximum_ROI_STAT			JointVariance_GLCM	X	X	maximum_ROI_STAT - ADC		
Range_STAT		X	JointEntropy_GLCM	X	X	Range_STAT - ADC		X
x1stPercentile_STAT		X	diffAverage_GLCM	X	X	x1stPercentile_STAT - ADC		X
x10thPercentile_STAT			diffVariance_GLCM			x10thPercentile_STAT - ADC		
x25thPercentile_STAT			diffEntropy_GLCM			x25thPercentile_STAT - ADC		
x50thPercentile_STAT	X		sumAverage_GLCM			x50thPercentile_STAT - ADC	X	
x75thPercentile_STAT		X	sumVariance_GLCM	X	X	x75thPercentile_STAT - ADC		X
x90thPercentile_STAT			sumEntropy_GLCM	X		x90thPercentile_STAT - ADC		
x95thPercentile_STAT	X	X	angularSecondMoment_GLCM			x95thPercentile_STAT - ADC	X	X
IQR_STAT	X	X	contrast_GLCM			IQR_STAT - ADC	X	X
skewness_STAT	X	X	dissimilarity_GLCM			skewness_STAT - ADC	X	X
kurtosis_STAT	X	X	InverseDifference_GLCM			kurtosis_STAT - ADC	X	X
IntensityKurtosis_STAT			NormalisedInverseDifference_GLCM	X		IntensityKurtosis_STAT - ADC		
IntensityVariance_STAT		X	InverseDifferenceMoment_GLCM	X		IntensityVariance_STAT - ADC		X
mean_intensity_IH			NormalisedInverseDifferenceMoment_GLCM	X		mean_intensity_IH - ADC		
JointMax_GLCM		X	inverseVariance_GLCM	X		JointMax_GLCM - ADC		X
JointAverage_GLCM	X	X	correlation_GLCM	X	X	JointAverage_GLCM - ADC	X	X
JointVariance_GLCM	X	X	Autocorrelation_GLCM	X		JointVariance_GLCM - ADC	X	X
JointEntropy_GLCM	X		clustertendency_GLCM			JointEntropy_GLCM - ADC		
diffAverage_GLCM	X	X	clustershad_GLCM	X	X	diffAverage_GLCM - ADC	X	X
diffVariance_GLCM	X	X	clusterprominence_GLCM	X	X	diffVariance_GLCM - ADC	X	X
diffEntropy_GLCM	X		infCorr1_GLCM	X	X	diffEntropy_GLCM - ADC	X	
sumAverage_GLCM			infCorr2_GLCM	X		sumAverage_GLCM - ADC		
sumVariance_GLCM	X		SRE_GLRLM	X	X	sumVariance_GLCM - ADC	X	
sumEntropy_GLCM	X		LRE_GLRLM	X	X	sumEntropy_GLCM - ADC	X	
angularSecondMoment_GLCM	X	X	LGRE_GLRLM	X	X	angularSecondMoment_GLCM - ADC	X	X
contrast_GLCM		X	HGRE_GLRLM	X	X	contrast_GLCM - ADC		X
dissimilarity_GLCM			SRLGE_GLRLM			dissimilarity_GLCM - ADC		
InverseDifference_GLCM	X		SRHGE_GLRLM			InverseDifference_GLCM - ADC	X	
NormalisedInverseDifference_GLCM		X	LRLGE_GLRLM			NormalisedInverseDifference_GLCM - ADC		X
InverseDifferenceMoment_GLCM		X	LRHGE_GLRLM			InverseDifferenceMoment_GLCM - ADC		X
NormalisedInverseDifferenceMoment_GLCM			GLNU_GLRLM	X	X	NormalisedInverseDifferenceMoment_GLCM - ADC		
inverseVariance_GLCM			GLNU_norm_GLRLM			inverseVariance_GLCM - ADC		
correlation_GLCM	X		RLNU_GLRLM	X		correlation_GLCM - ADC	X	
Autocorrelation_GLCM		X	RLNU_norm_GLRLM			Autocorrelation_GLCM - ADC		X
clustertendency_GLCM			RP_GLRLM			clustertendency_GLCM - ADC		
clustershad_GLCM		X	GreylevelVariance_GLRLM	X		clustershad_GLCM - ADC		X
clusterprominence_GLCM			RunlengthVariance_GLRLM			clusterprominence_GLCM - ADC		
infCorr1_GLCM	X	X	RunEntropy_GLRLM	X		infCorr1_GLCM - ADC	X	X
infCorr2_GLCM		X				infCorr2_GLCM - ADC		X
SRE_GLRLM						SRE_GLRLM - ADC		
LRE_GLRLM	X					LRE_GLRLM - ADC	X	
LGRE_GLRLM	X	X				LGRE_GLRLM - ADC	X	X
HGRE_GLRLM	X	X				HGRE_GLRLM - ADC	X	X
SRLGE_GLRLM		X				SRLGE_GLRLM - ADC		X
SRHGE_GLRLM						SRHGE_GLRLM - ADC		
LRLGE_GLRLM						LRLGE_GLRLM - ADC		
LRHGE_GLRLM		X				LRHGE_GLRLM - ADC		X
GLNU_GLRLM	X					GLNU_GLRLM - ADC	X	
GLNU_norm_GLRLM						GLNU_norm_GLRLM - ADC		
RLNU_GLRLM						RLNU_GLRLM - ADC		
RLNU_norm_GLRLM	X					RLNU_norm_GLRLM - ADC	X	
RP_GLRLM		X				RP_GLRLM - ADC		X
GreylevelVariance_GLRLM						GreylevelVariance_GLRLM - ADC		
RunlengthVariance_GLRLM		X				RunlengthVariance_GLRLM - ADC		X
RunEntropy_GLRLM	X					RunEntropy_GLRLM - ADC	X	
						JointMax_GLCM - T2		
						JointAverage_GLCM - T2	X	X
						JointVariance_GLCM - T2	X	X
						JointEntropy_GLCM - T2	X	X
						diffAverage_GLCM - T2	X	X
						diffVariance_GLCM - T2		X
						diffEntropy_GLCM - T2		
						sumAverage_GLCM - T2		
						sumVariance_GLCM - T2	X	X
						sumEntropy_GLCM - T2	X	
						angularSecondMoment_GLCM - T2		
						contrast_GLCM - T2	X	X
						dissimilarity_GLCM - T2		
						InverseDifference_GLCM - T2		
						NormalisedInverseDifference_GLCM - T2		X
						InverseDifferenceMoment_GLCM - T2		X
						NormalisedInverseDifferenceMoment_GLCM - T2		
						inverseVariance_GLCM - T2	X	
						correlation_GLCM - T2	X	X
						Autocorrelation_GLCM - T2	X	
						clustertendency_GLCM - T2		
						clustershad_GLCM - T2	X	X
						clusterprominence_GLCM - T2	X	X
						infCorr1_GLCM - T2	X	X
						infCorr2_GLCM - T2		X
						SRE_GLRLM - T2	X	X
						LRE_GLRLM - T2	X	X
						LGRE_GLRLM - T2	X	X
						HGRE_GLRLM - T2	X	X
						SRLGE_GLRLM - T2		
						SRHGE_GLRLM - T2		
						LRLGE_GLRLM - T2		
						LRHGE_GLRLM - T2		
						GLNU_GLRLM - T2	X	X
						GLNU_norm_GLRLM - T2		X
						RLNU_GLRLM - T2	X	X
						RLNU_norm_GLRLM - T2		
						RP_GLRLM - T2		
						GreylevelVariance_GLRLM - T2		
						RunlengthVariance_GLRLM - T2		
						RunEntropy_GLRLM - T2	X	

Table 2.14: Features selected (marked with X) by AP algorithm.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.8: 7-fold crossvalidation svm performances with polynomial kernel, in terms of accuracy, sensitivity, specificity, of training set (blue line) and test set (red line), for 3D and 2D datasets, with Gaussian filter ($\sigma = 0.5mm$) and 0.5 mm interpolation. Only features with AUC value greater than 0.7 are considered.

figure 2.9, 2.10 and 2.11, respectively. As done before, accuracy, sensitivity and specificity are shown. In addition, the optimized threshold for classifying lesions to one or the other class is shown.

As mentioned in section 2.4.2, the redundant features, highly correlated to others already present, have been eliminated. In the table 2.16 the number of features remaining after applying the three correlation limit thresholds (0.99, 0.98, 0.95) is shown.

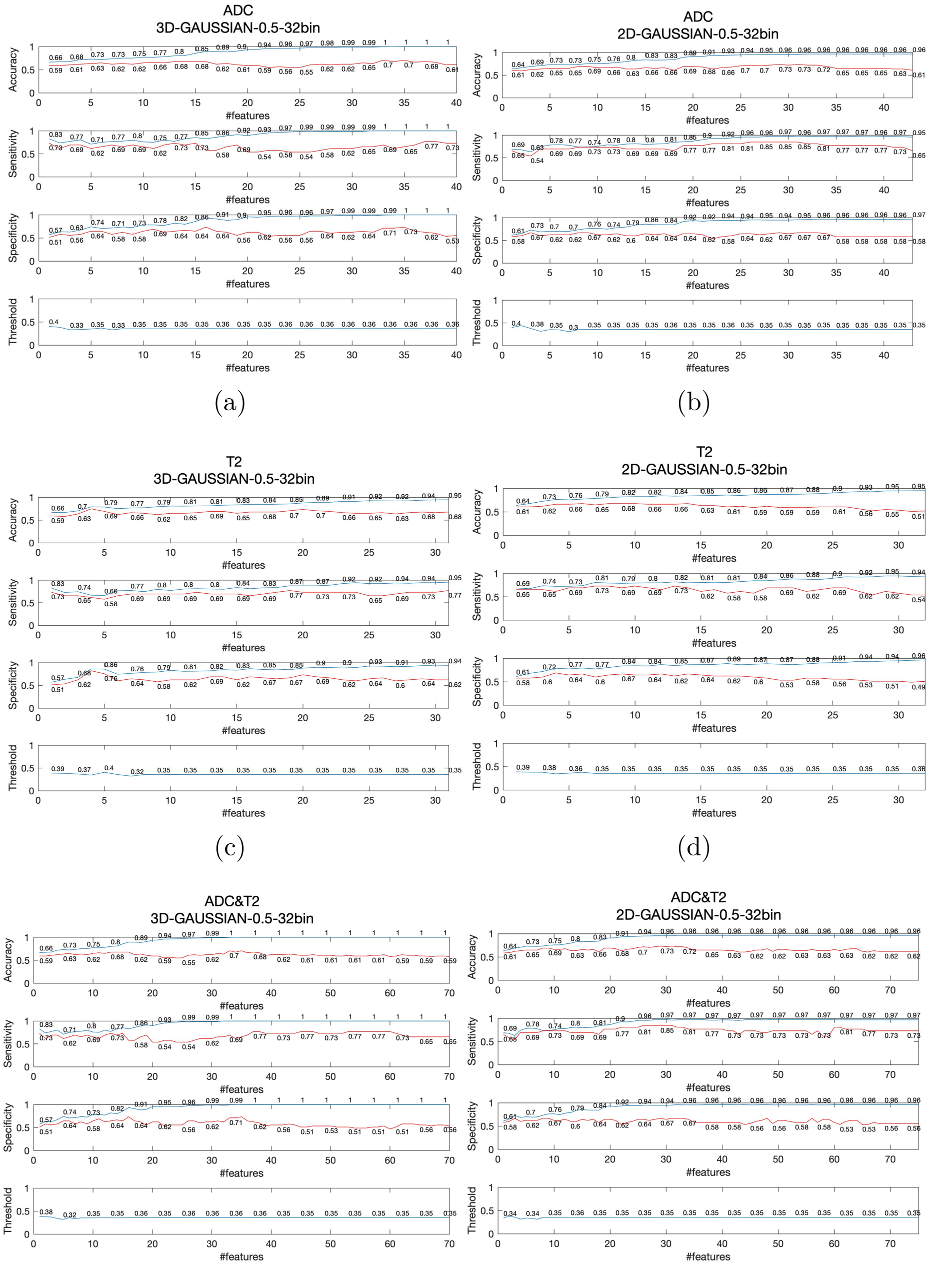
		ADC 3D - 2D	T2 3D - 2D	ADC-T2 joined 3D - 2D
	0.99	40 - 43	31 - 32	70 - 75
Correlation threshold	0.98	35 - 38	25 - 29	59 - 66
	0.95	26 - 31	17 - 23	42 - 53

Table 2.16: Number of remaining features in 3D and 2D datasets with Gaussian filter ($\sigma = 0.5mm$)and 0.5 mm interpolation, after eliminating highly correlated features: for each pair of features with correlation greater than or equal to a set threshold (0.99, 0.98 and 0.95), the feature with the lowest AUC is deleted.

Observing the performance trends obtained in the training (blue line) and in the test (red line) set in the figure 2.9, 2.10 and 2.11, the aim is to find a bifurcation point where the performances of the training set continue to increase while those of the test set flatten or decrease. This results in the search for the overfitting point. In the table 2.17 the cut-offs identified are shown, in terms of number of features. In addition, the optimization of the classification threshold shows that the best threshold is always less than 50% (default threshold). It varies in the range between 24% and 49%, but mainly tends to 35%, 33% and 38% in the cases of svm classifiers with polynomial kernel, svm with Gaussian kernel, and RF classifier, respectively.

3rd FS method: MRMR

Figure 2.12 shows the results obtained by training an svm classifier with polynomial and Gaussian kernels with the features selected by the MRMR algorithm (see section 2.5.2). Model performances are shown in terms of True Positives (TP) and True Negatives (TN) on training (b) and test (d) set, and are directly compared with the total number of low-aggressive (0) and high-aggressive (1) lesions, on training (a) and test (c) sets. In this way it is visibly immediate to compare the results between the different models, and to compare each model to the distribution of the two classes, i.e. ideal performance with 100% accuracy.

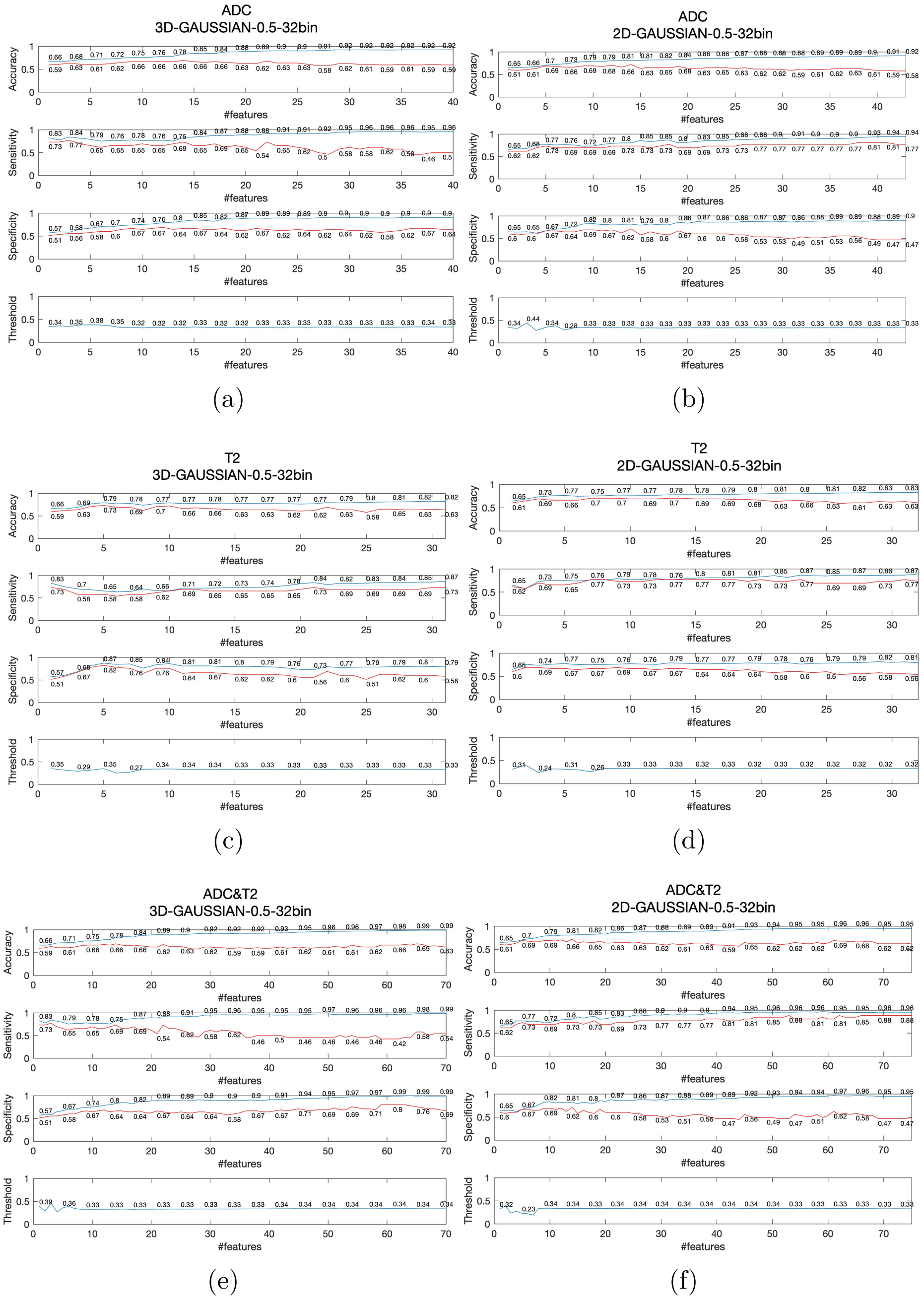


(a) (b)

(c) (d)

(e) (f)

Figure 2.9: 7-fold crossvalidation svm performances with polynomial kernel, in terms of accuracy, sensitivity, specificity, of training set (blue line) and test set (red line), for 3D and 2D datasets, with Gaussian filter ($\sigma = 0.5mm$) and 0.5 mm interpolation. All features are considered.

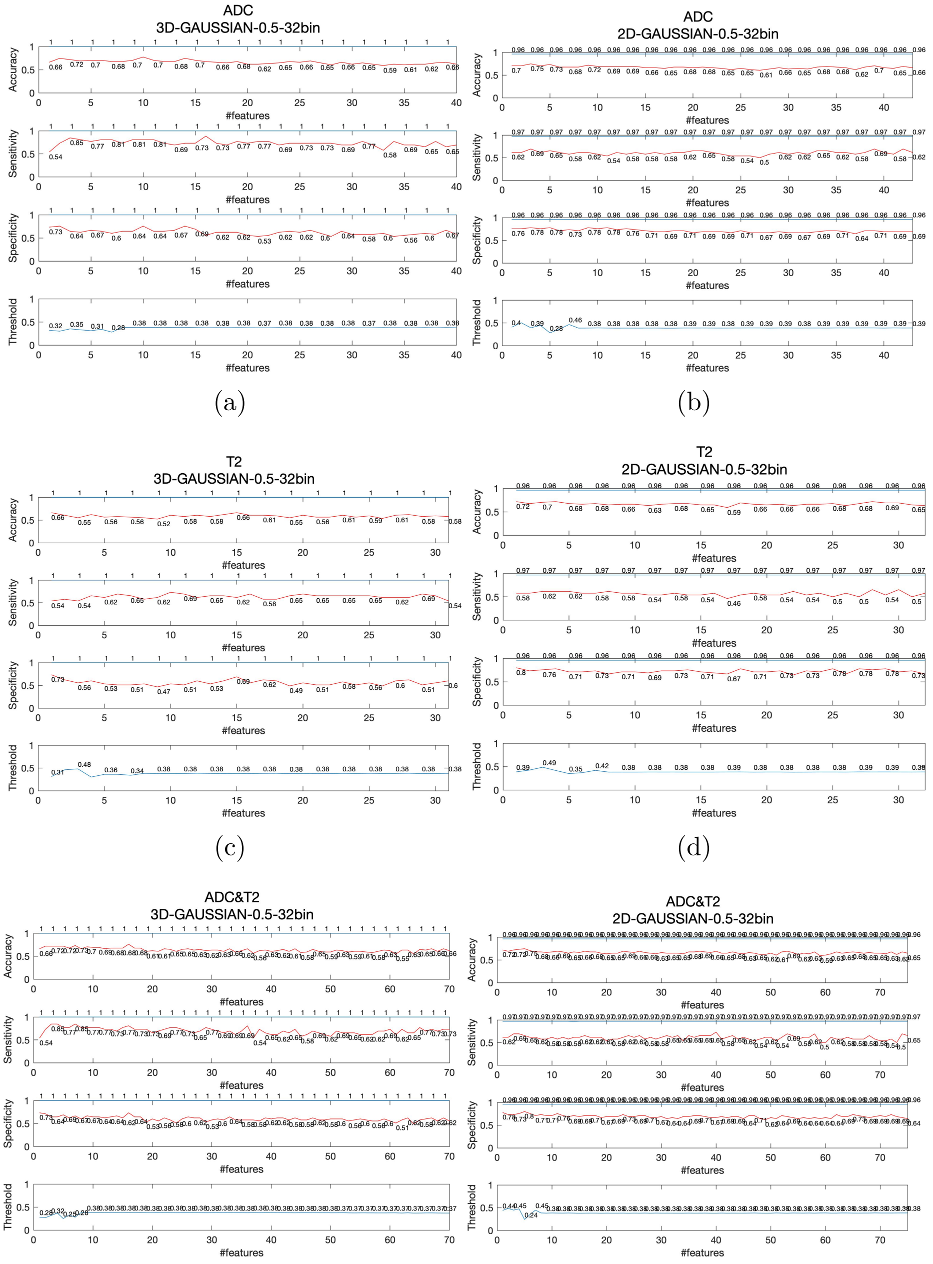


(a) (b)

(c) (d)

(e) (f)

Figure 2.10: 7-fold crossvalidation svm performances with Gaussian kernel, in terms of accuracy, sensitivity, specificity, of training set (blue line) and test set (red line), for 3D and 2D datasets, with Gaussian filter ($\sigma = 0.5mm$) and 0.5 mm interpolation. All features are considered.



(a)

(b)

(c)

(d)

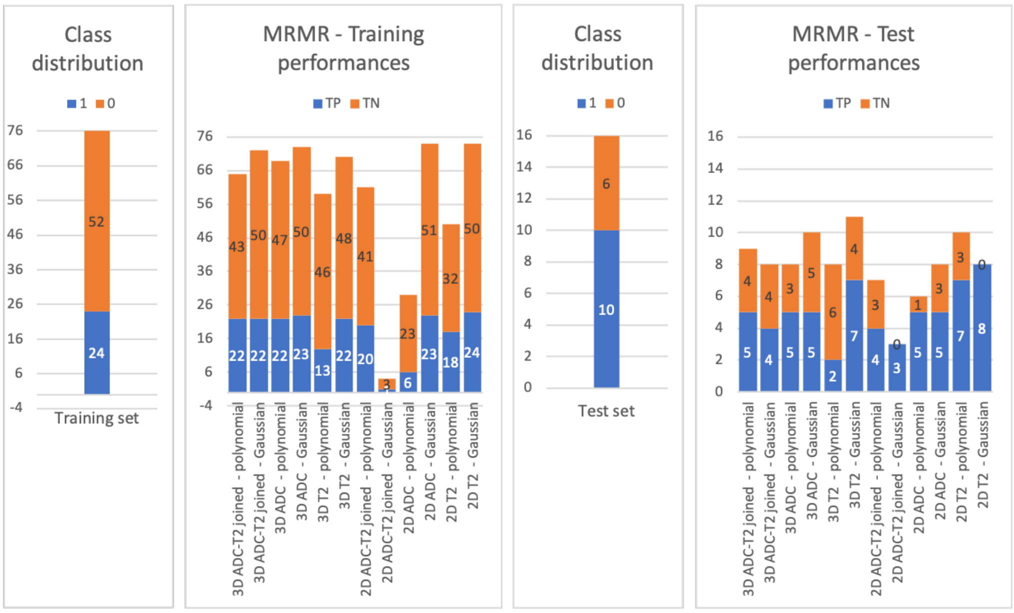
(e)

(f)

Figure 2.11: 7-fold crossvalidation random forest performances with 100 trees, in terms of accuracy, sensitivity, specificity, of training set (blue line) and test set (red line), for 3D and 2D datasets, with Gaussian filter ($\sigma = 0.5mm$)and 0.5 mm interpolation. All features are considered.

	3D		2D	
	Polynomial	Gaussian	Polynomial	Gaussian
ADC	4	5	9	7
T2	15	14	10	5
ADC-T2 joined	15	12	11	10

Table 2.17: Number of features corresponding to the overfitting cut-off identified by adding the features in descending order of AUC value and observing the performance in training and test sets. The cut-off point was identified by analyzing the performances obtained by training a svm classifier with polynomial and Gaussian kernel, with ADC, T2 and ADC-T2 joined datasets, both in the 3D and 2D case.



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

Figure 2.12: Results of svm models trained with polynomial and Gaussian kernel, and with features selected by Minimum Redundance Maximum Relevance (MRMR) algorithm. Specifically, the distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a) and test (c) set is shown. Bar diagrams in (b) and (d) show the performances in terms of number of True Positives (TP) (blue bars) and True Negatives (TN) (orange bars), in training and test set, respectively. Results for ADC, T2 and ADC-T2 joined datasets, in the cases of 3D and 2D features with Gaussian filter ($\sigma = 0.5mm$) and interpolation at 0.5mm.

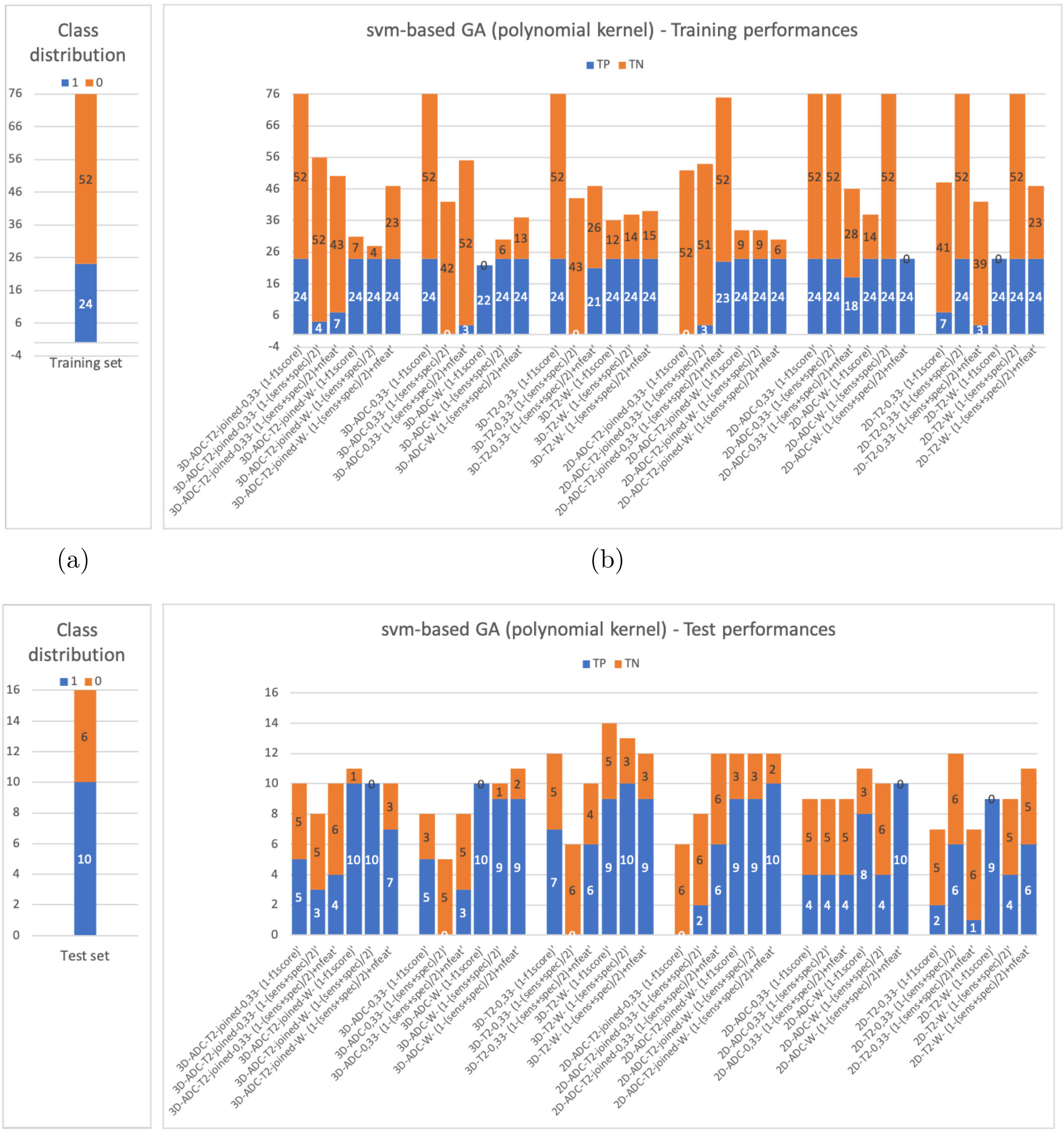
Comparing the results obtained on the training set (figure 2.12 (b)) the best model is that of the svm classifier trained with Gaussian kernel and with 2D T2 dataset, obtaining only two False Positive (FP). On the other hand, the worst model is that of the classifier trained with a Gaussian kernel and 2D ADC-T2 joined dataset, obtaining an accuracy of 5,3%. However, the best model on the training set, described before, does not achieve good performances with the test set: it misclassifies only 2 low-aggressive and 1 high-aggressive lesion.

4th FS method: GA

Figures 2.13 and 2.14 show the results obtained by training an svm classifier respectively with polynomial and Gaussian kernels, in training (a) and test set (b), with the features selected by the GA algorithm (see section 2.5.2).
Observing the models obtained with features selected with the svm-based GA, we note how the use of the Gaussian kernel allows to obtain higher training results than the polynomial kernel. The results obtained in the test set are instead comparable. Also in the case of FS through MIQ-based GA (figure 2.15, the resulting models have better performance with the Gaussian kernel than with the polynomial. Specifically, this is true for the training of all models and for the testing of almost all models (only the 3D dataset T2 performs best with polynomial kernel).

5th FS method: AP

Figures 2.16 shows the results obtained by training an svm classifier with polynomial and Gaussian kernels, in training (a) and test set (b), with the features selected by the AP algorithm (see section 2.5.2). Observing the performance on the training set, all svm models trained with polynomial kernel achieve 100% accuracy. The remaining models correctly classify correctly almost all lesions, misclassifying a maximum of 3 lesions. Moving on to the results on the test set, the model trained with 3D T2 dataset and polynomial kernel reaches the highest accuracy, equal to 75%.



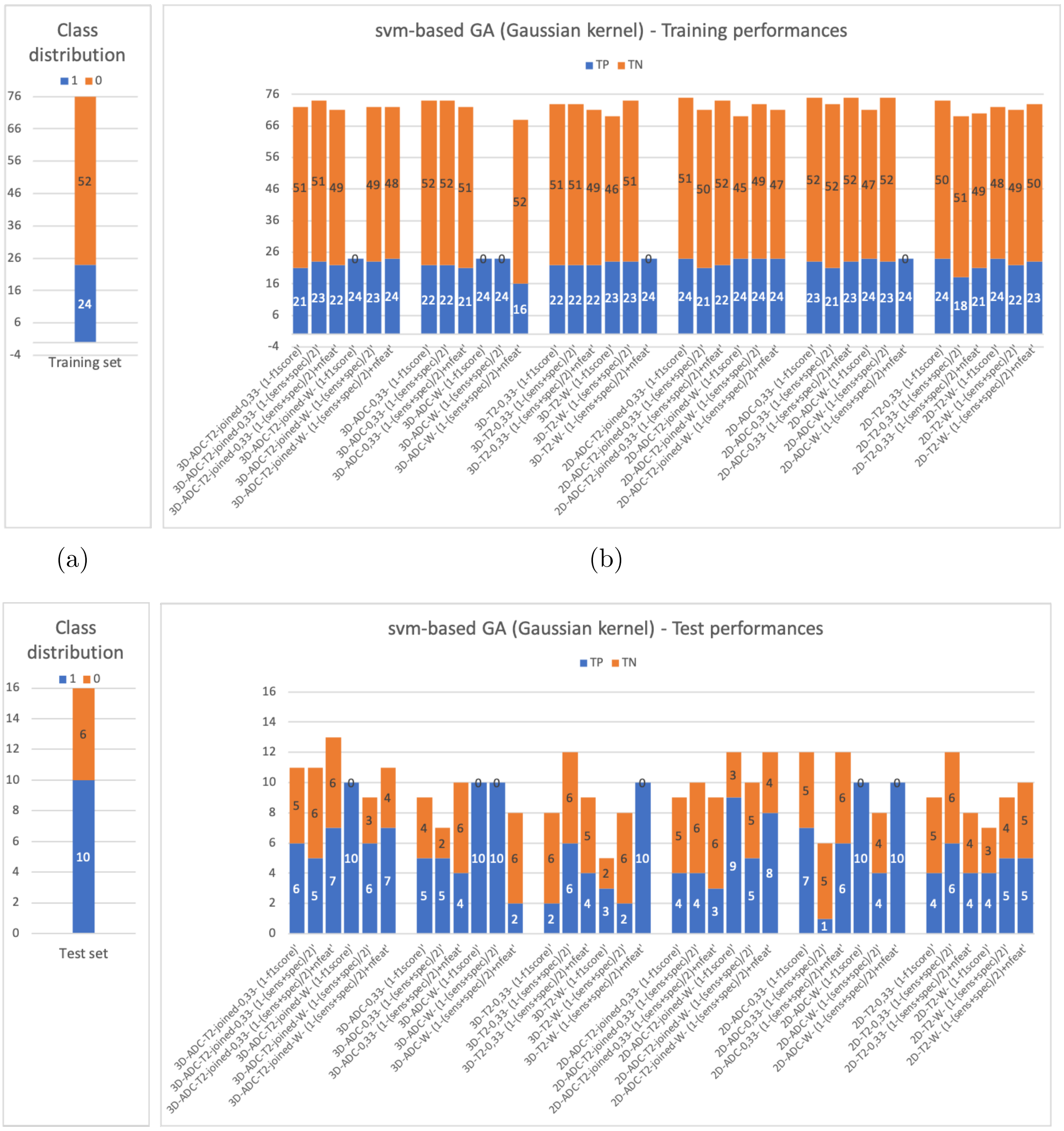
(a)

(b)

(c)

(d)

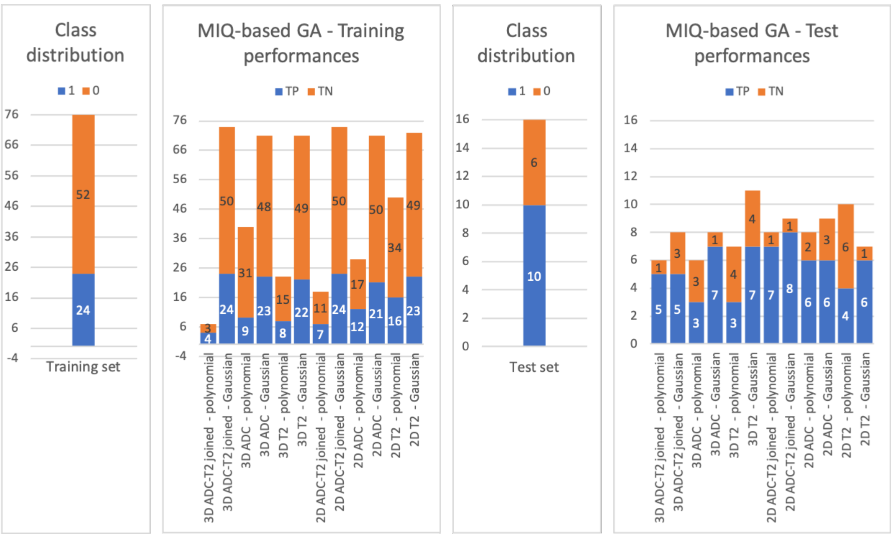
Figure 2.13: Results of svm models trained with polynomial kernel, and with features selected by Genetic Algorithm (GA). Specifically, the distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a) and test (c) set is shown. Bar diagrams in (b) and (d) show the performances in terms of number of True Positives (TP) (blue bars) and True Negatives (TN) (orange bars), in training and test set, respectively. Results for ADC, T2 and ADC-T2 joined datasets, in the cases of 3D and 2D features with Gaussian filter ($\sigma = 0.5mm$) and interpolation at 0.5mm.



(a) (b)

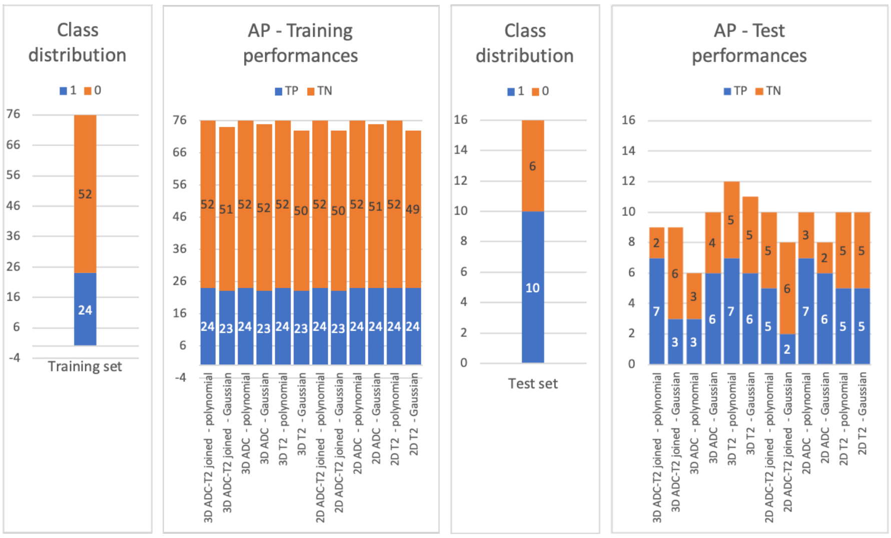
(c) (d)

Figure 2.14: Results of svm models trained with Gaussian kernel, and with features selected by Genetic Algorithm (GA). Specifically, the distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a) and test (c) set is shown. Bar diagrams in (b) and (d) show the performances in terms of number of True Positives (TP) (blue bars) and True Negatives (TN) (orange bars), in training and test set, respectively. Results for ADC, T2 and ADC-T2 joined datasets, in the cases of 3D and 2D features with Gaussian filter ($\sigma = 0.5mm$) and interpolation at 0.5mm.



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

Figure 2.15: Results of svm models trained with polynomial and Gaussian kernel, and with features selected by MIQ-based Genetic Algorithm (GA). Specifically, the distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a) and test (c) set is shown. Bar diagrams in (b) and (d) show the performances in terms of number of True Positives (TP) (blue bars) and True Negatives (TN) (orange bars), in training and test set, respectively. Results for ADC, T2 and ADC-T2 joined datasets, in the cases of 3D and 2D features with Gaussian filter ($\sigma = 0.5mm$) and interpolation at 0.5mm.



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

Figure 2.16: Results of svm models trained with Gaussian kernel, and with features selected by Affinity Propagation (AP) algorithm. Specifically, the distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a) and test (c) set is shown. Bar diagrams in (b) and (d) show the performances in terms of number of True Positives (TP) (blue bars) and True Negatives (TN) (orange bars), in training and test set, respectively. Results for ADC, T2 and ADC-T2 joined datasets, in the cases of 3D and 2D features with Gaussian filter ($\sigma = 0.5mm$) and interpolation at 0.5mm.

2.5.4 Validation

To understand the generalization capability of the models, the best classifier of each FS method is selected and used to predict the lesion class of the validation set. Figure 2.17 shows the performances of these models in training (b), test (d) and validation (f). The results are shown, as in the previous sections, in terms of number of TPs and TNs. Table 2.18 shows the corresponding performance indices. Note that these indices are very sensitive to the cardinality of the subset, especially when the latter contains few samples. For example, in the case of the validation set, the presence of only one False Positive causes the specificity to drop from 100% to 75%. For this reason, it is relevant to evaluate the results also in terms of number of TPs and TNs.

		training			test			validation	
	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity
AP	1,00	1,00	1,00	0,75	0,70	0,83	0,64	0,56	1,00
AUC	1,00	1,00	1,00	0,69	0,70	0,67	0,68	0,67	0,75
MIQ-based GA	0,97	1,00	0,96	0,56	0,80	0,17	0,82	0,83	0,75
svm-based GA	1,00	1,00	1,00	0,75	0,70	0,83	0,59	0,56	0,75
MRMR	0,97	0,96	0,98	0,50	0,50	0,50	0,82	0,83	0,75

Table 2.18: Performance indices of the best ML classifiers, in training, test, and validation set.

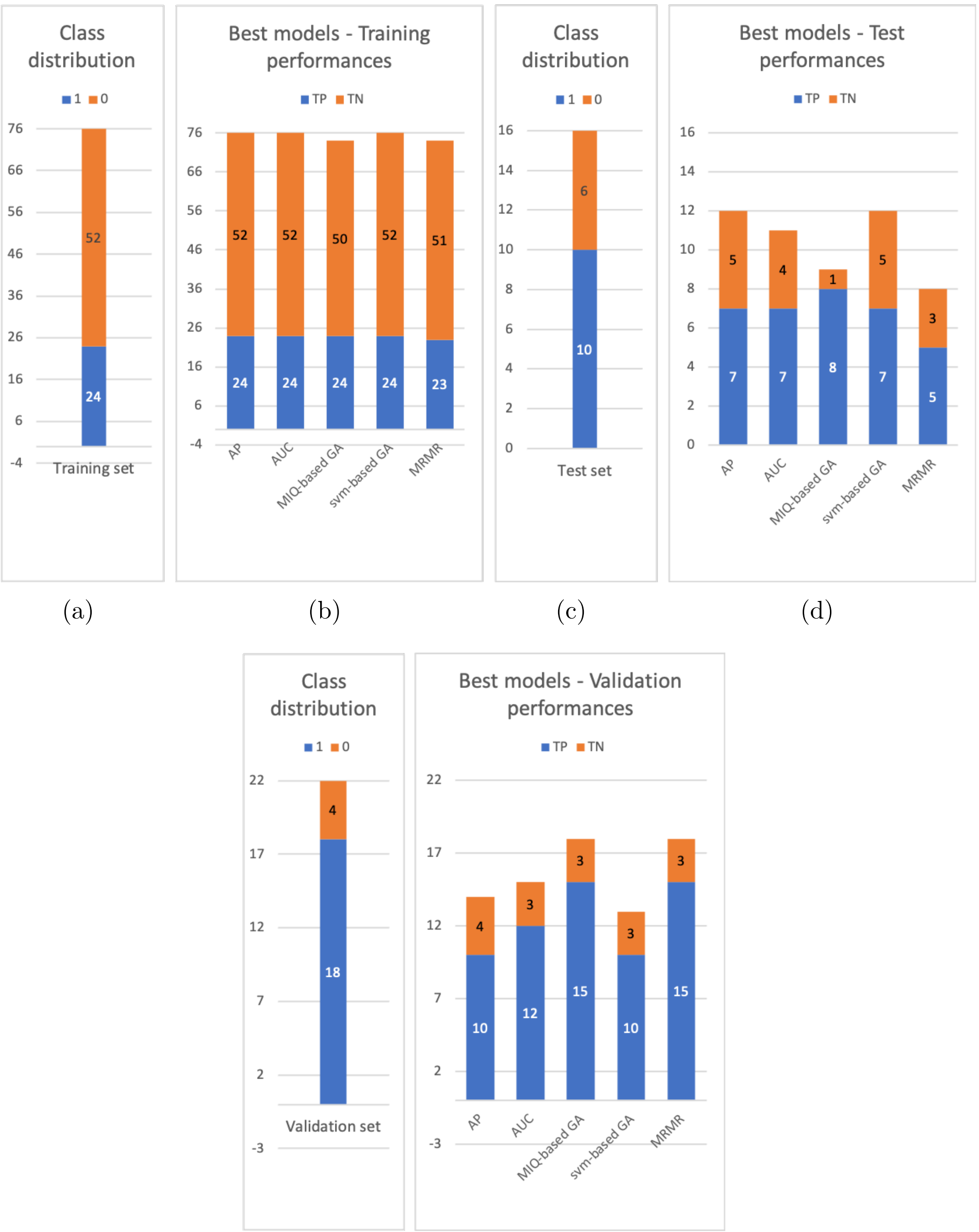
2.6 Discussion

The Machine learning approach started by evaluating the twenty datasets extracted from the ADC maps and T2W images. The univariate analysis allowed to evaluate the ability of each feature to predict the classification of lesions. Specifically, this was assessed on the entire dataset from Candiolo. In particular, the non-parametric analysis of the Mann-Whitney U test and the calculation of the feature-output correlation coefficient did not lead to relevant results. On the contrary, the analysis of the AUC of each feature showed a greater predictivity of the 3D and 2D datasets with interpolation at 0.5mm and Gaussian filter ($sigma = 0.5mm$) compared to the others. Specifically, 3D achieved a greater number of relevant features (with AUC> 70%). While, considering only the first four features in descending order of AUC, 2D obtained the lowest number of misclassified lesions. Of these two configurations, the performances of 32 and 64 bins were equivalent. Therefore, from here on, only the 3D and 2D datasets, with interpolation at 0.5mm, and Gaussian filter ($sigma = 0.5mm$) have been analyzed, both for ADC maps and T2W images.

Subsequently, on these selected datasets, a multiparametric analysis was carried out, using three different algorithms: AP, GA, MRMR. Of the three methods, the MRMR selected the subsets with the fewest features, with up to five features selected. In contrast, the AP has always selected at least the forty percent of the features of each dataset.

However, in order to evaluate the performance of such FS methods, it is necessary to observe the performance of the resulting models. Looking at figure 2.17, it can be seen that two models have two misclassified and the remaining three reach 100% accuracy with the training set. This would suggest a possible overfitting of the models. Then, observing the performances on the test set, two models, deriving from AP and svm-based GA methods, obtain the best performances, misclassifying only three high-aggressive lesions (FN) and one low-aggressive (FP). At the same time, however, these two models are the two that perform worse with the validation set. Conversely, the two models that achieve the worst results with the test set

(MIQ-based GA and MRMR) achieve the highest accuracy with the validation set. To better understand these results, it is necessary to remember that the test set consists entirely of Candiolo lesions and that the validation set, on the contrary, contains only Molinette lesions. Therefore, we must consider the difference between the images of the two hospitals, variability that may depend on the hand of the clinician who performed the MRI examination and on the different machines used for the scan, as well as consider the intrinsic variability of the type of tumor.



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

(e) (f)

Figure 2.17: Results in terms of True Positive (TP) and True Negative (TN) of the best models for each type of Feature Selection method, in training (b), test (d), and validation (f) set. The distribution of the two classes (0 low-aggressive and 1 high-aggressive tumors) in training (a), test (c) and validation (e) set is shown.

Chapter 3

Deep Learning for Prostate Cancer Aggressiveness characterization

3.1 Introduction to Deep Learning

Deep learning is a branch of machine learning and it is based on artificial Neural Networks (aNN). As the name suggests, aNN algorithms are born with the aim of mimicking the human cognitive process. They are composed of a series of computational units, the Neurons, called Perceptrons, which exchange information through synapses. The connection between two neurons is characterized by a weight that represents the strength of communication between those two neurons, and must be optimized during the learning process. Figure 3.1 shows the analogy between neurons and perceptrons.

Like in the cerebral cortex, perceptrons are interconnected in Layers. Specifically, an aNN is composed of:

- Input layer, that represents the input as a fixed-length vector of numbers;
- Output layer, that provides the output of the aNN;
- Hidden layer, that stands between two layers and which output can be an input for another hidden layer or for an output layer (in any case not an output of the system).

Deep Neural Networks (DNN) are deep as they are composed of multiple hidden layers. Unlike machine learning algorithms, which require feature extraction and selection phases, deep learning algorithms are able to extract high dimensional

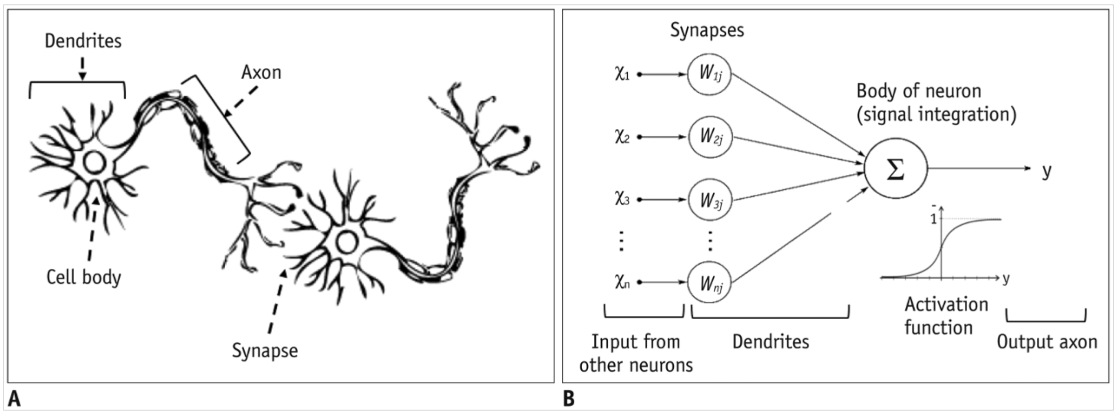
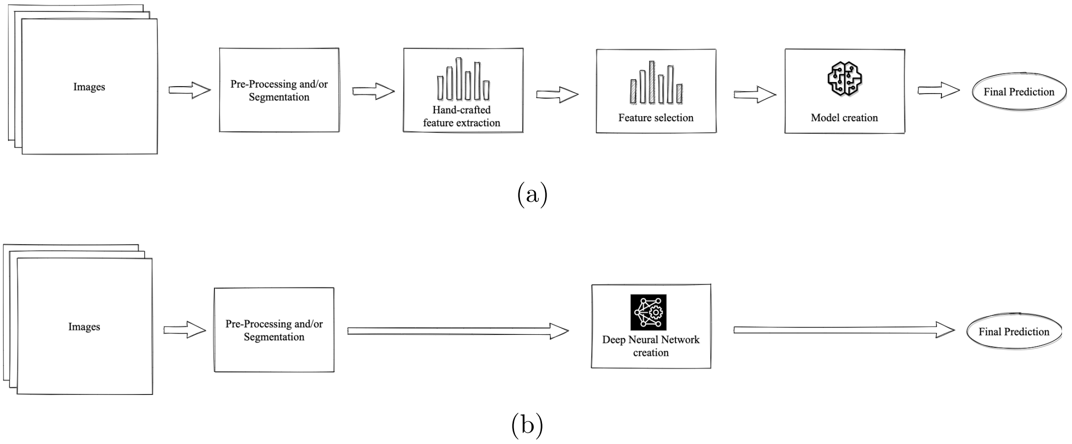


Figure 3.1: Conceptual analogy between real neurons (A) and artificial neurons (B) [34].

features. Starting from the extraction of simple local features, the level of abstraction is gradually increased by moving forward with the hidden layers. This process avoids the use of hand-crafted features, a time-consuming process, and allows the use of more complex sets of features than traditional machine learning ones. Moreover, DNN is able to perform a huge number of routine, repetitive tasks, within a relatively shorter period of time. Figure 3.2 shows the two different pipelines associated with machine learning (b) and deep learning (c) techniques.


Images
Pre-Processing and/or Segmentation
Hand-crafted feature extraction
Feature selection
Model creation
Final Prediction
Images
Pre-Processing and/or Segmentation
Deep Neural Network creation
Final Prediction

(a)

(b)

Figure 3.2: Different pipelines associated with machine learning (b) and deep learning (c) techniques.

3.1.1 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a type of Deep Neural Network optimized for processing images. The following sections describe the main CNN layers and their characteristics.

Convolutional Layer

Convolutional Layer is the first layer of a CNN and it is used for feature extraction. It is named in this way because it applies sliding convolutional filters to the input image. Specifically, the convolution process consists of the sum of multiplication products between the starting image and a matrix of numbers, called Kernel, centered in a pixel of the image, to which the obtained value is associated. This operation is performed repeatedly displacing and centering the kernel in all the pixels of the image. At the end a matrix, called Feature Map, is obtained, composed of the values coming from the convolution process (see figure 3.4 A).
A problem that arises when carrying out the convolution process is that of overlapping the outermost pixels of the image with the kernel. Generally, to avoid excluding those pixels, which would result in a decrease in the size of the feature map, zero padding is performed. This technique involves adding columns and rows of zeros to the matrix of the original image, so as to be able to center the kernel even in the most extreme pixels and perform the convolution in all the pixels of the image. This results in a feature map of the same size as the input image.
The type of pattern that can be detected depends on the type and size of the kernel used. Moreover, a convolutional layer is generally composed of multiple kernels, or filters, the number of which determines the Depth of the convolutional layer.
Some characteristics of the Layer are:

- It is connected only with the previous layer. Computationally, the convolution becomes very efficient, as each neuron is connected to a limited number of inputs.

- A neuron is sensitive to the inputs of only neighboring neurons of the previous layer. This mimics the behavior of the visual cortex which is divided into areas containing neurons, each specialized in a specific task. In this way, different populations of neural cells are sensitive to different levels of visual patterns.

- Various neurons in different areas share the same connection and therefore the same weight. These neurons will have the same type of sensitivity but for different visual areas. In this way, the number of parameters that the network will have to optimize during the training phase are reduced.

Non-Linear Layer

Non-Linear Layer is is located immediately after a convolutional layer and it consists in the use of a non-linear activation function. This allows the network to process more advanced, non-linear data. The most widely used are Rectified Linear Unit (ReLU), Sigmoid and Tanh.

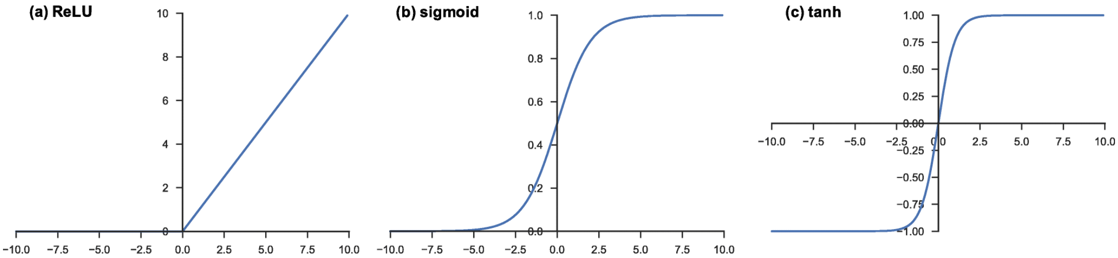
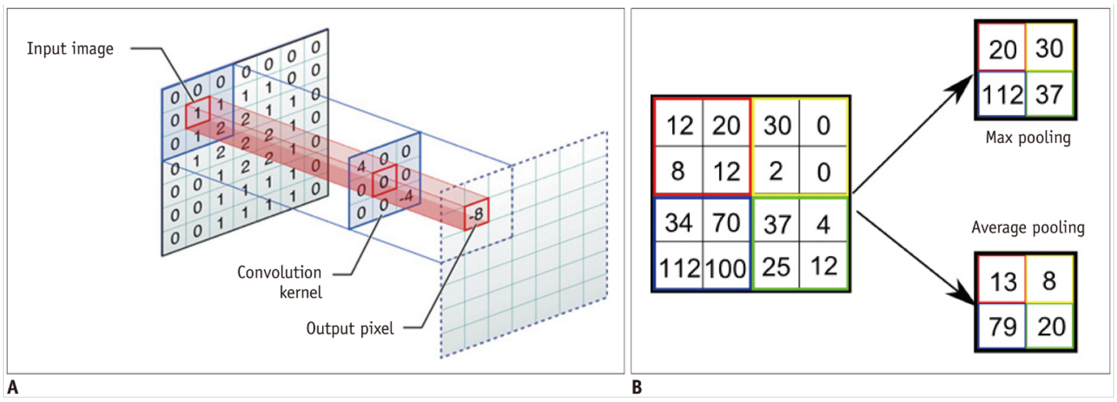


Figure 3.3: Activation functions commonly applied to neural networks: (a) rectified linear unit (ReLU), (b) sigmoid, and (c) hyperbolic tangent (tanh) [35].

Pooling Layer

Pooling Layer performs a feature reduction operation. The simplest way is by dividing the features into distinct rectangles and for each of them selecting an element, which can be the maximum value (Max Pooling) or the average value (Average pooling) (see figure 3.4 B).



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

13	8
79	20

Figure 3.4: Illustration of convolution (A) and pooling methods (B) [34].

Fully Connected Layer

Fully Connected (FC) Layer has the task of predicting the final output class, starting from the image patterns identified by the previous layers. As with the convolutional layer, the fully connected layer is followed by a non-linear layer. The last layer of the network generally uses the so-called Softmax activation function since it will allow us to direct the output into a class.

Figure 3.5 shows a basic structure of convolution neural network for binary classification.

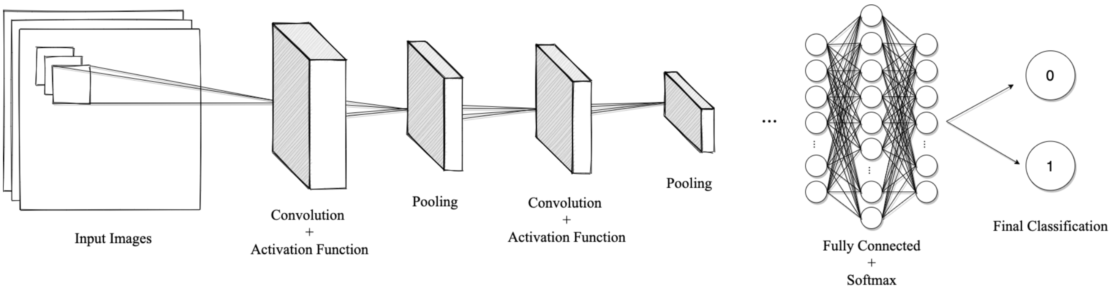


Figure 3.5: Basic structure of a convolution neural network for binary classification.

3.2 Dataset

3.2.1 Patients

The patients used to build Deep Learning classifiers are the same as those used in the previous chapter on Machine Learning. Table 3.1 summarizes the number of patients, the number of lesions and the number of MRI slices (the tumor can be present in multiple MRI slices), from each hospital.

	Patients	Lesions	MRI slices
Candiolo IRCCS	58	73	145
San Giovanni Molinette Hospital	43	43	170

Table 3.1: Number of patients, lesions and MRI slices coming from the two hospitals, Candiolo IRCCS and San Giovanni Molinette Hospital.

3.2.2 Dataset creation

In the case of deep learning algorithms, as mentioned in the introductory section 3.1, the input data of the network are directly the images of the lesions. Due to the small number of patients available, it was decided to extract square patches, or ROIs, of side 3 and 5 pixels in 2-pixel steps, from each image (see figure 3.6), to use as input for CNNs. Specifically, only ROIs completely inside the segmented lesion are selected, both from T2W images and ADC maps.

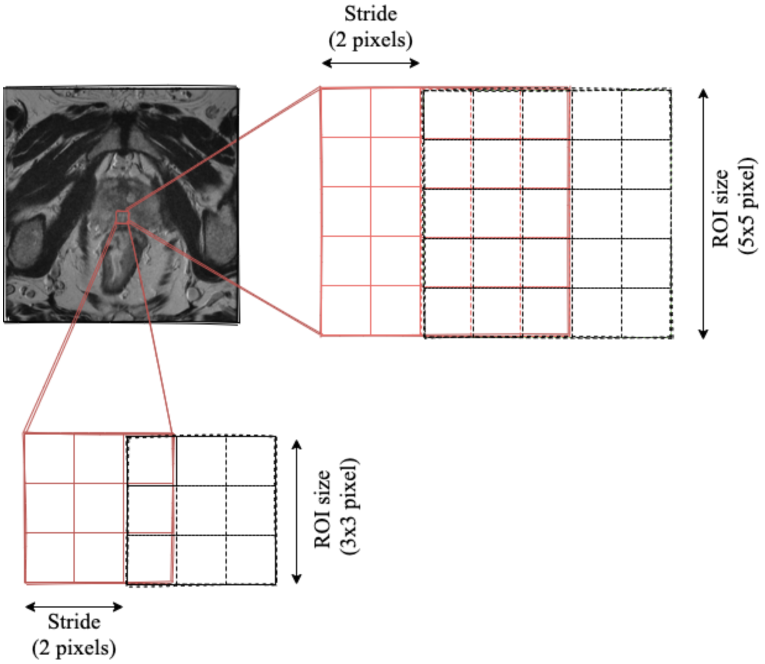


Figure 3.6: Extraction of ROIs, 3x3 and 5x5, with stride of 2 pixels from a T2W image.

Note that extrapolation of 5x5 ROIs was not possible in all MRI slices: some MRI slices contained such a small lesion portion that they could not totally contain a 5x5 ROI. For this reason, one low-aggressive lesion coming from a patient of Candiolo IRCCS is not present, since the tumor is too small in all MRI slices.

3.3 Classifier construction

Several CNN structures are tested, different for the number of Convolutional and FC Layers, for the number and size of the filters, and for the number of neurons per

FC Layer. However, the CNN structure always follows the same pattern: an input layer; one, two, or three Convolutional Layers (each followed by a Relu Layer); one or two FC Layers (the last always consisting of two neurons); a final Softmax Layer that provides the final classification. Specifically, it was decided not to use Pooling Layer because ROIs used as CNN input have a very small size, therefore it is not necessary to reduce the size of the resulting feature maps.
Tables 3.2, 3.3, and 3.4 show the number of filters for each Convolutional Layer and the number of neurons for each FC Layer in the cases of structures with 1, 2, and 3 convolutional layers, respectively. Table 3.5, instead, shows filter sizes set for these CNN structures. Note that in the case of 3x3 ROIs, structures have a maximum of two Convolutional Layers, and have only filters with 3x3 dimensions. Instead, in the case of 5x5 ROIs, structures with three Convolutional Layers and filters with 5x5 dimensions are also tested.

	Configurations					
	#1	#2	#3	#4	#5	#6
Input layer	-	-	-	-	-	-
2D Convolutional Layer + ReLu	5	10	15	20	25	30
Fully Connected Layer	6	10	15	20	25	25
Fully Connected layer + Softmax	2	2	2	2	2	2

Table 3.2: CNN structures tested with 1 Convolutional Layer. Number of filters of the convolutional layer and number of neurons for each Fully Connected Layer.

	Configurations					
	#1	#2	#3	#4	#5	#6
Input layer	-	-	-	-	-	-
(1st) 2D Convolutional Layer + ReLu	5	10	15	20	25	30
(2nd) 2D Convolutional Layer + ReLu	7	15	20	25	30	40
Fully Connected Layer	6	10	15	20	25	25
Fully Connected layer + Softmax	2	2	2	2	2	2

Table 3.3: CNN structures tested with 2 Convolutional Layers. Number of filters per each convolutional layer and number of neurons for each Fully Connected Layer.

The parameters set for all CNN structures are shown in the table 3.6, the other parameter values are set by default by the Matlab routine.

The output of all CNN structures is the prediction of the ROI class. Once that the predictions of all ROIs have been obtained, two steps are necessary to predict

	Configurations				
	#1	#2	#3	#4	#5
Input layer	-	-	-	-	-
(1st) 2D Convolutional Layer + ReLu	5	10	15	20	25
(2nd) 2D Convolutional Layer + ReLu	7	15	20	25	30
(3rd) 2D Convolutional Layer + ReLu	9	20	25	30	35
Fully Connected Layer	6	10	15	20	25
Fully Connected layer + Softmax	2	2	2	2	2

Table 3.4: CNN structures tested with 3 Convolutional Layers. Number of filters per each convolutional layer and number of neurons for each Fully Connected Layer.

ROI size	1 ConvLayer	2 ConvLayers		3 ConvLayers		
(pixel)		1st	2nd	1st	2nd	3rd
3x3	3x3	3x3	3x3	-	-	-
5x5	3x3	3x3	3x3	3x3	3x3	3x3
	5x5	5x5	5x5	5x5	5x5	5x5
	-	3x3	5x5	-	-	-
	-	5x5	3x3	-	-	-

Table 3.5: Filter size in structures with 1, 2, or 3 convolutional layers, according to the two dimensions of the ROIs.

the lesion: from the prediction of the ROI to that of the MRI slice, and from the latter to that of the lesion. To do this, the following method is chosen:

1. Slice MRI is classified as high-aggressive (1) if the percentage of its ROIs classified as high-aggressive are greater than a threshold. This threshold is both set equal to 50% (majority voting) and optimized in order to obtain the best values of sensitivity and specificity.

2. The lesion is classified once taking into account all the slices and applying a threshold on them (50% and optimized, as done in point 1 with ROIs), and also taking into account the prediction obtained only for the slice containing the largest portion of the tumor.

In total, four different predictions are obtained for the same lesion. The flowchart in the figure 3.7 summarizes the steps from ROI to lesion prediction.

Parameter	Choice	**Meaning**
Solver	sgdm	Stochastic gradient descent with momentum
InitialLearnRate	0,01	How much to change the weights in response to the estimated error; if too low training will take a long time, if too high probable stuck at a suboptimal result
MaxEpochs	2	Number of complete passes through the training set
Shuffle	every-epoch	Data shuffled before every training epoch
MiniBatchSize	64	Number of samples used to update weights
ValidationFrequency	25	Number of iterations between evaluations of validation metrics

Table 3.6: Parameters set for CNN structures.

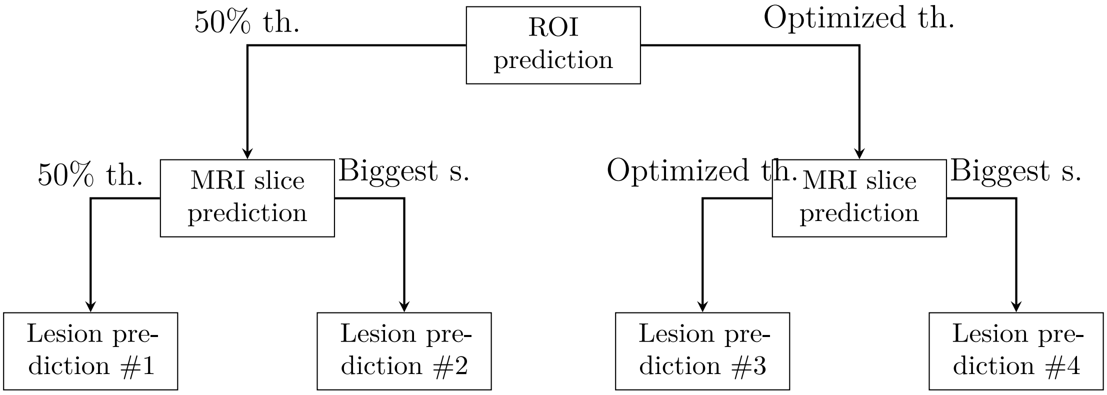


Figure 3.7: Flowchart from ROI to lesion prediction. 50% th=Majority voting; Optimized th.=Optimized threshold; Biggest=Biggest slice.

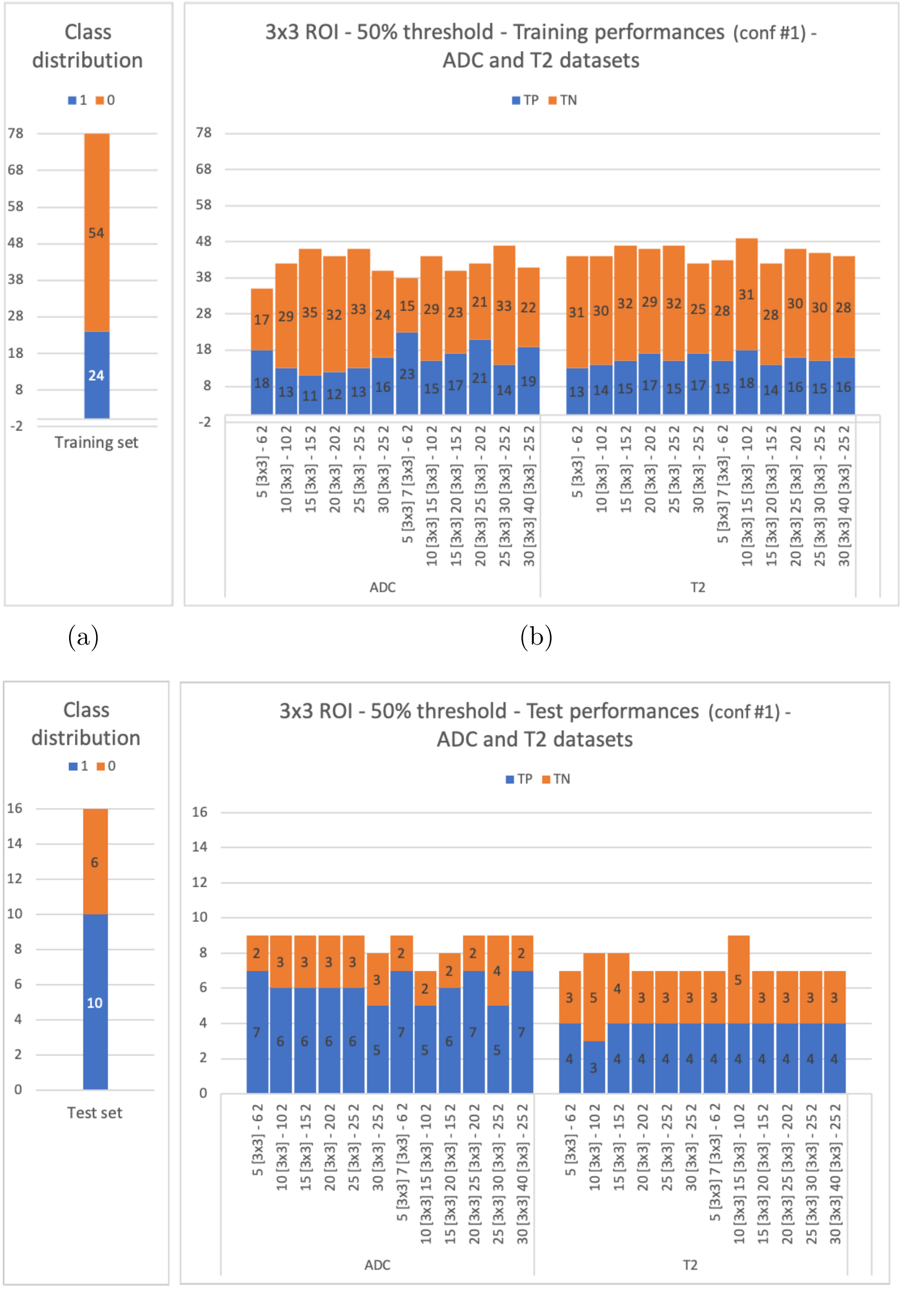
3.4 Results

3.4.1 All CNN structures: training and test set

The performances obtained for lesion prediction #1, #2 , #3, and #4, for both ADC and T2 datasets with 3x3 ROIs, are shown in figure 3.8, 3.9, 3.10 and 3.11, respectively.
The corresponding performances with 5x5 ROIs for lesion prediction #1, #2, #3, and #4 are shown in figure 3.14, 3.16 and 3.18, respectively, for ADC dataset; in figure 3.13, 3.15, 3.17 and 3.19, respectively, for T2 dataset.

Observing the performance of the 3x3 ROIs in all four types of prediction, the accuracy of the training set is around 60% and is approximately equivalent for the

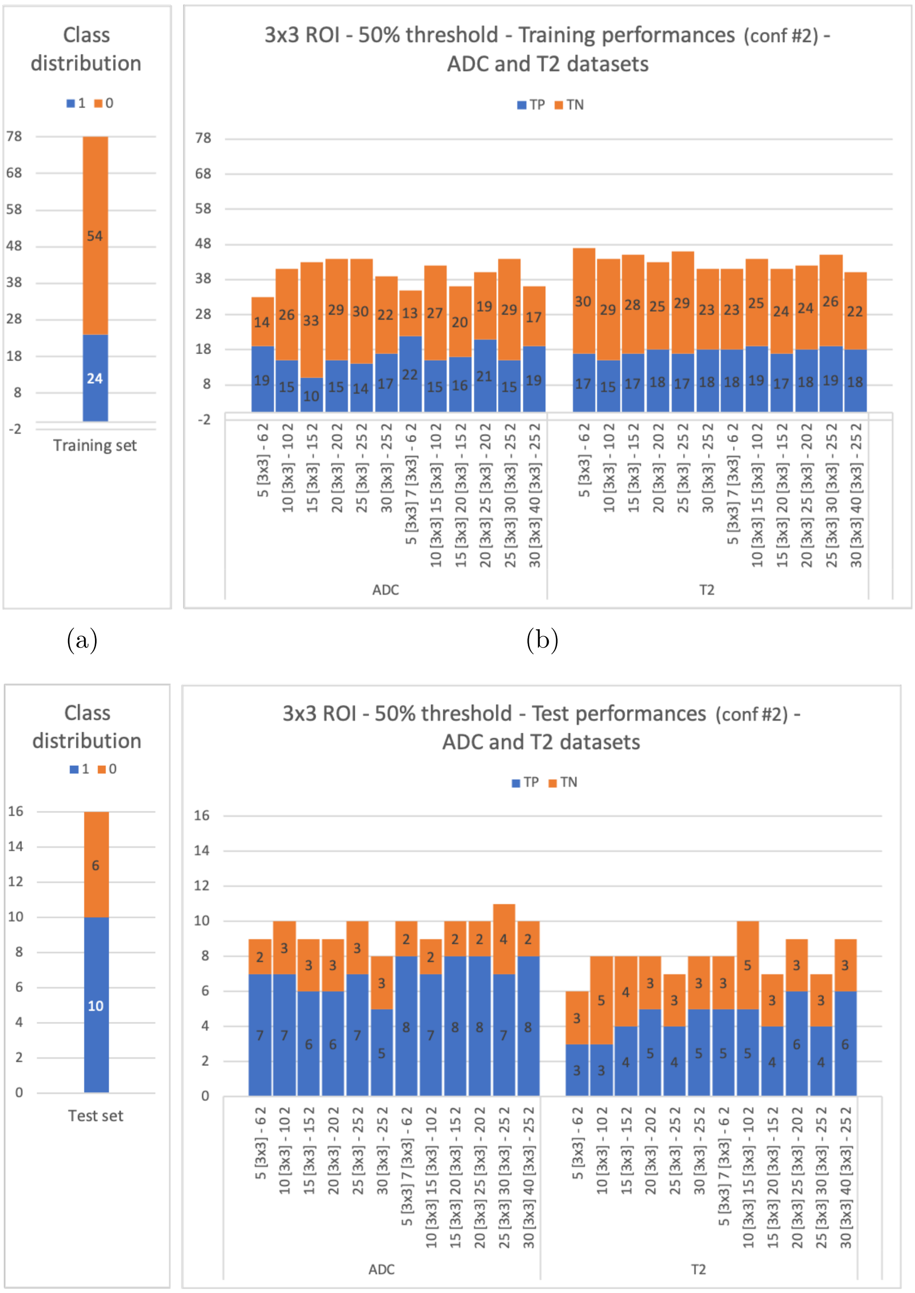
ADC and T2 datasets. Instead, observing the test set, the ADC dataset seems to be able to better predict the high-aggressive class, obtaining about 3 TP more than the ADC dataset.
Turning to the observation of CNNs with 5x5 ROI, it seems that the classifications of the lesions that take into account only the MRI slice containing the largest portion of the tumor (prediction #2 and #4) are able to correctly classify the low-aggressive class but not the high-aggressive one. In contrast, lesion classifications that apply a threshold on the predictions of MRI slices (predictions 1 and 3) have a high sensitivity but low specificity, which results in the classification of almost all the samples in the test set as high-aggressive.



(a) (b)

(c) (d)

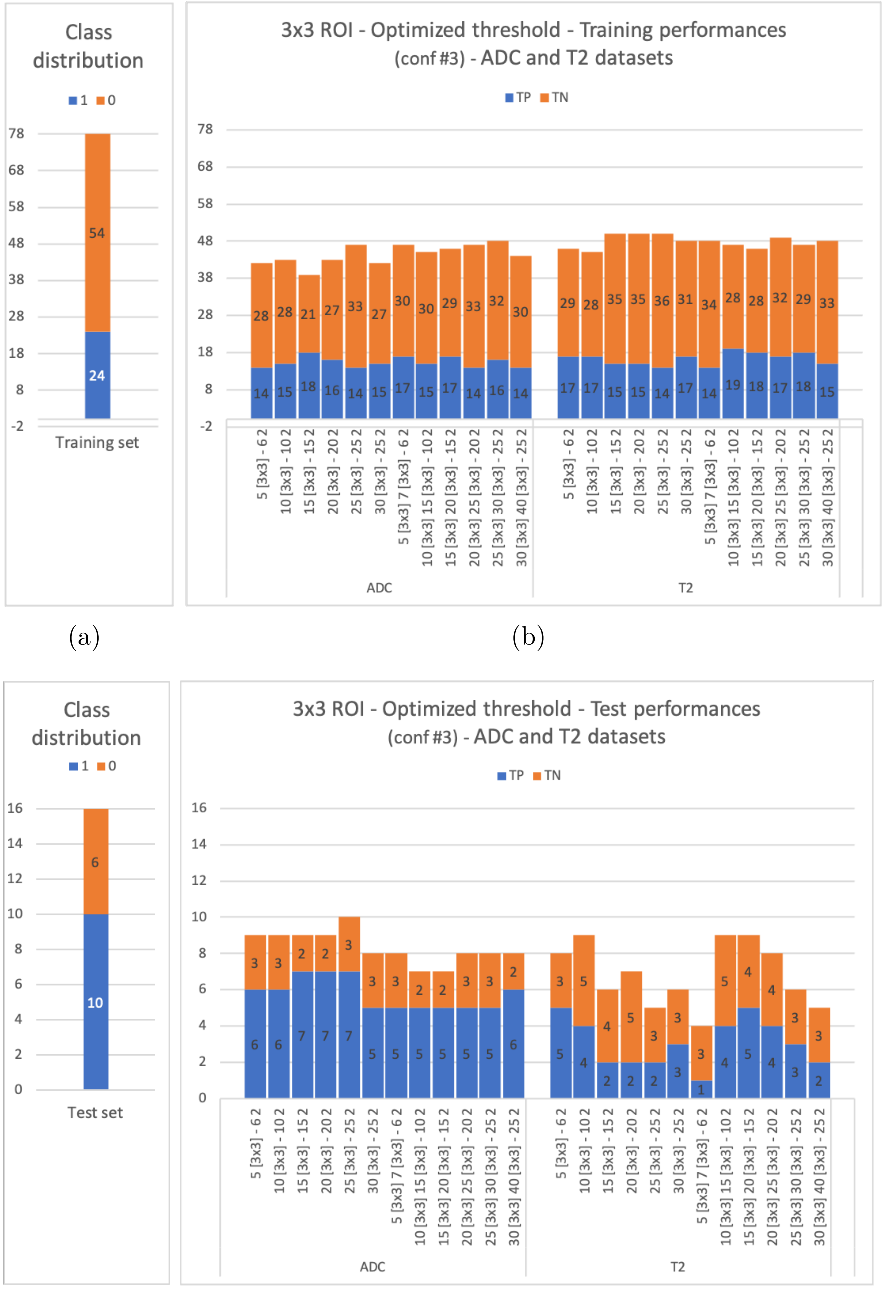
Figure 3.8: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 3x3 ROI and lesion prediction #1, i.e. with majority vote on ROIs and then on the MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

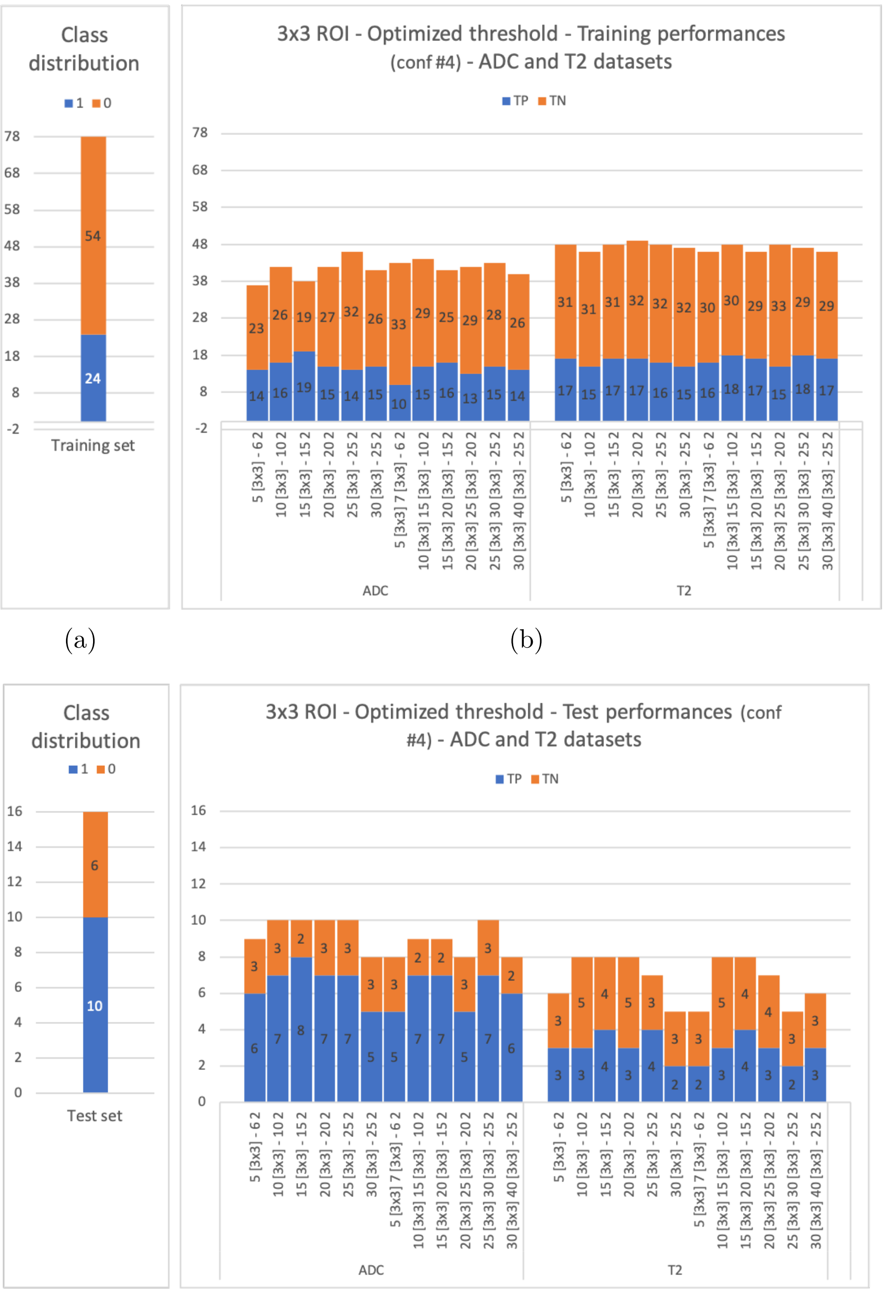
Figure 3.9: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 3x3 ROI and lesion prediction #2, i.e. with majority vote on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.10: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 3x3 ROI and lesion prediction #3, i.e. with optimed cut-off on ROIs and then on the MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



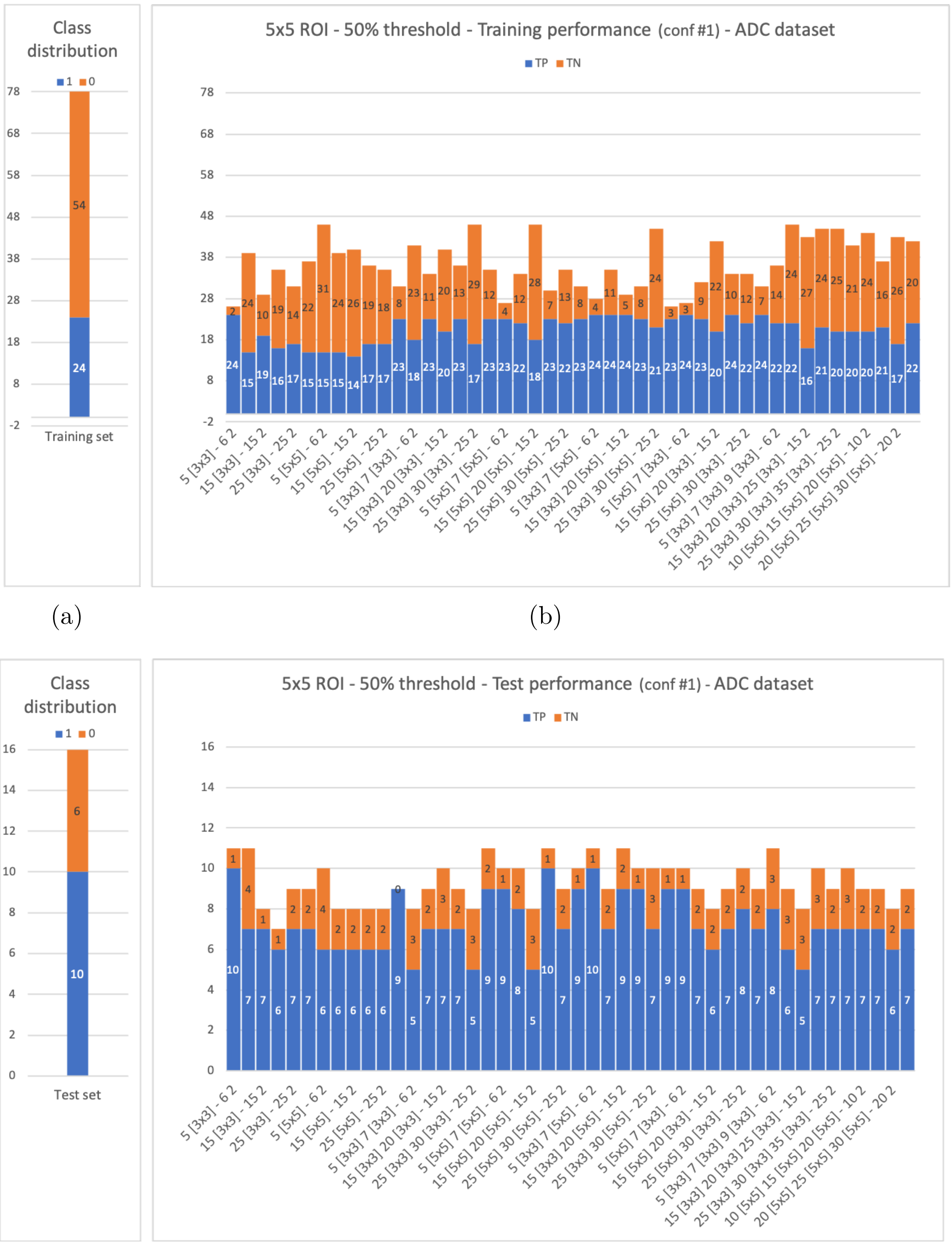
(a)

(b)

(c)

(d)

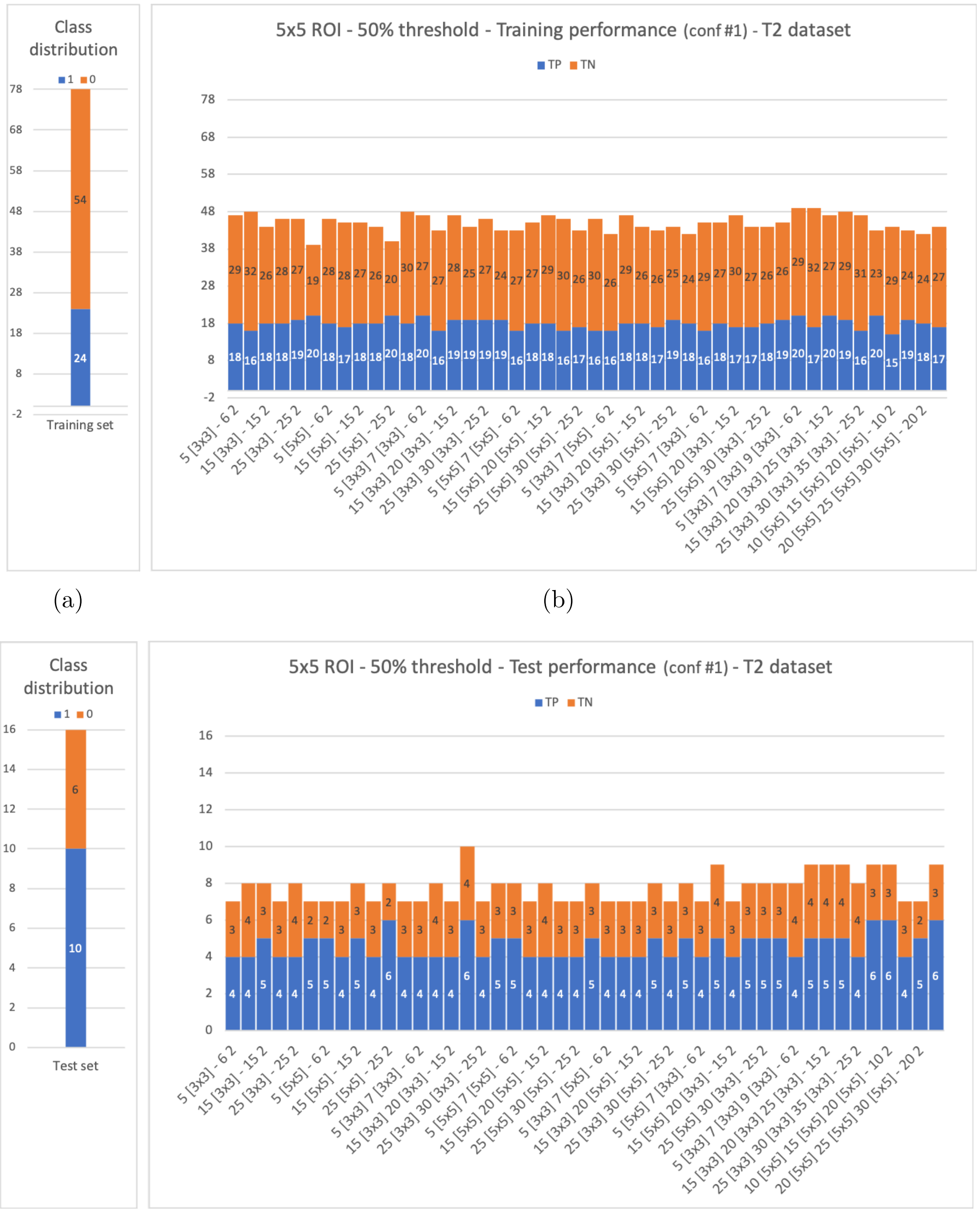
Figure 3.11: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 3x3 ROI and lesion prediction #4, i.e. with optimized cut-off on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.12: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, ADC dataset, and lesion prediction #1, i.e. with majority vote on ROIs and then on the MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.13: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, T2 dataset, and lesion prediction #1, i.e. i.e. with majority vote on ROIs and then on the MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.

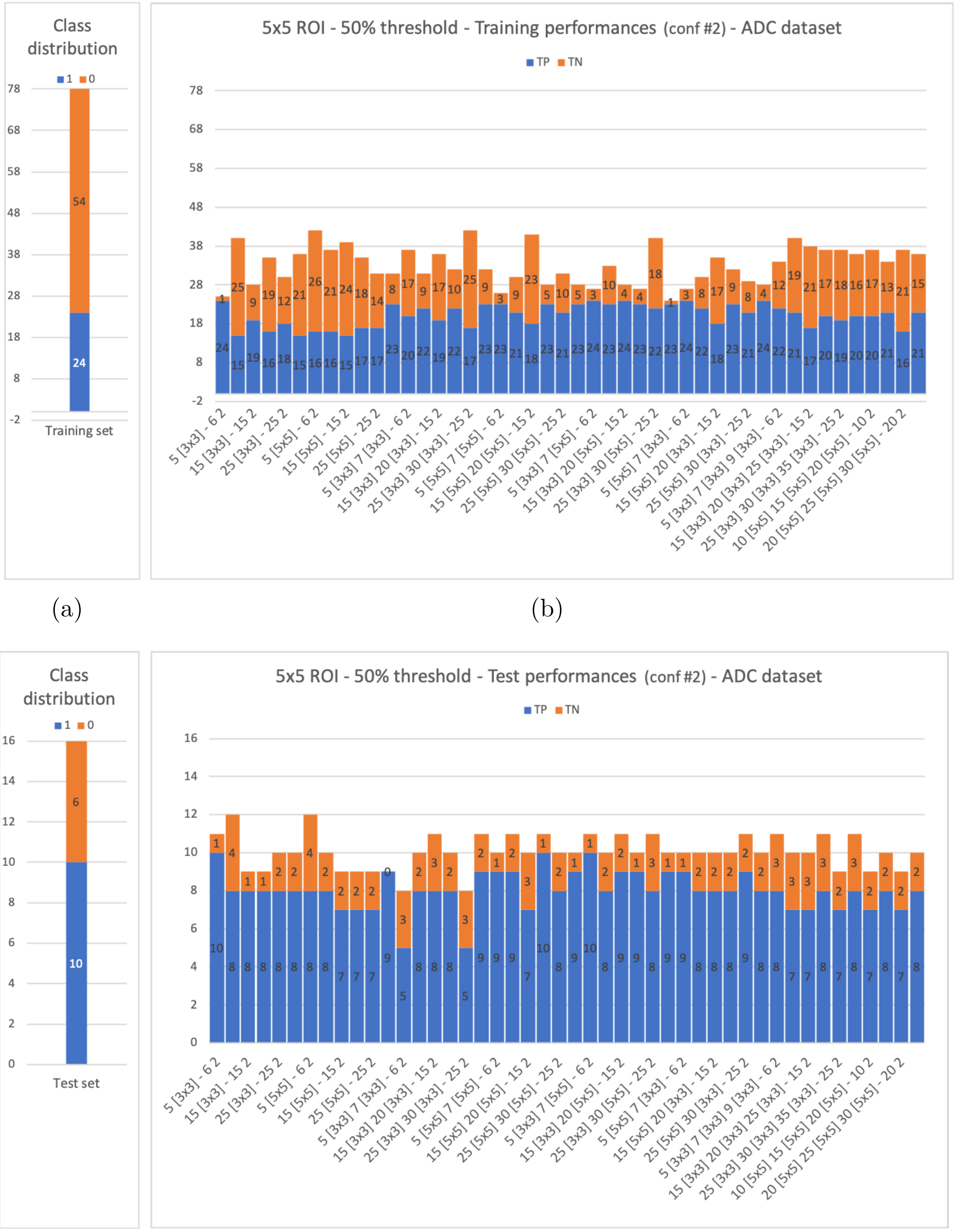
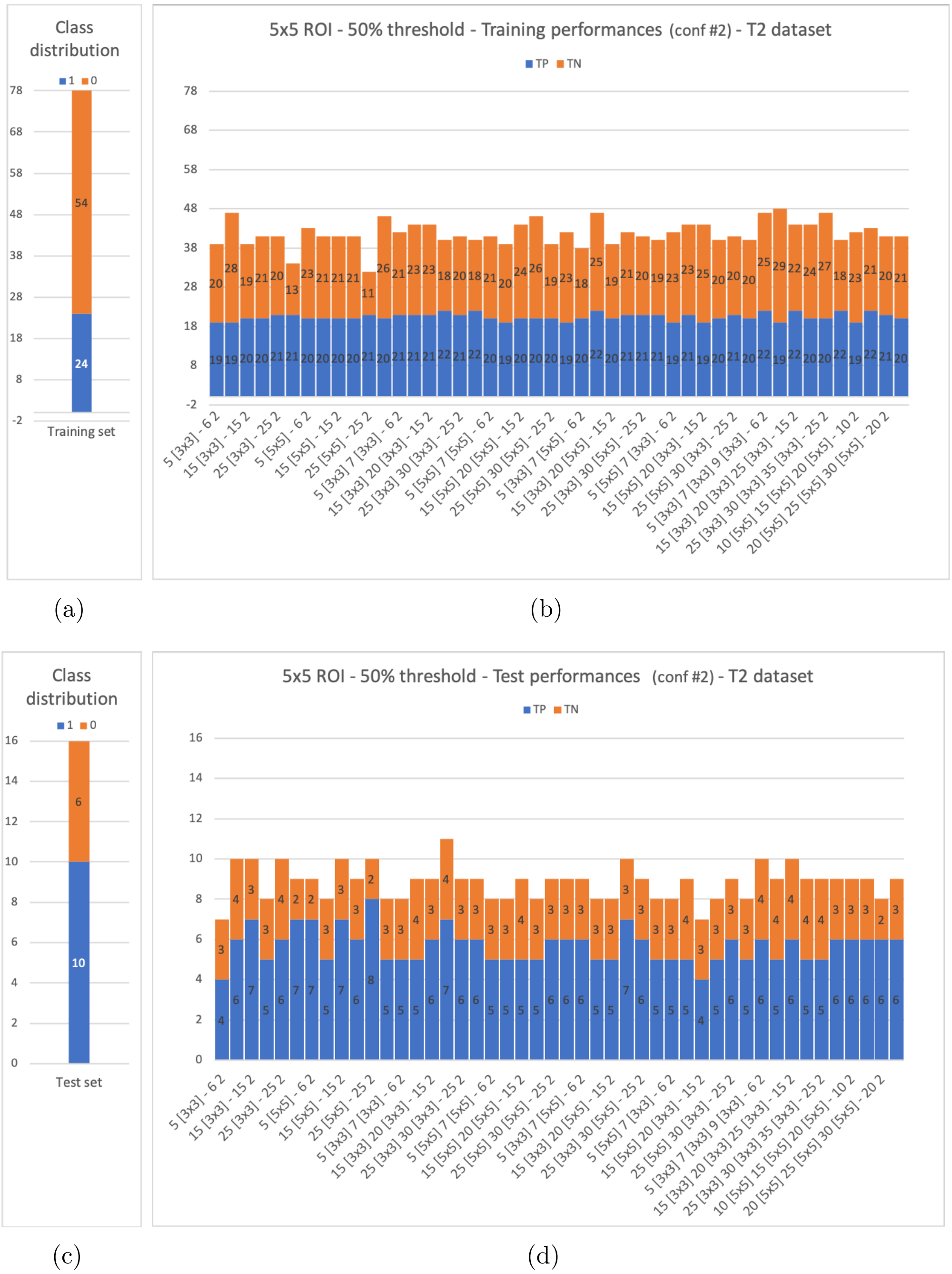


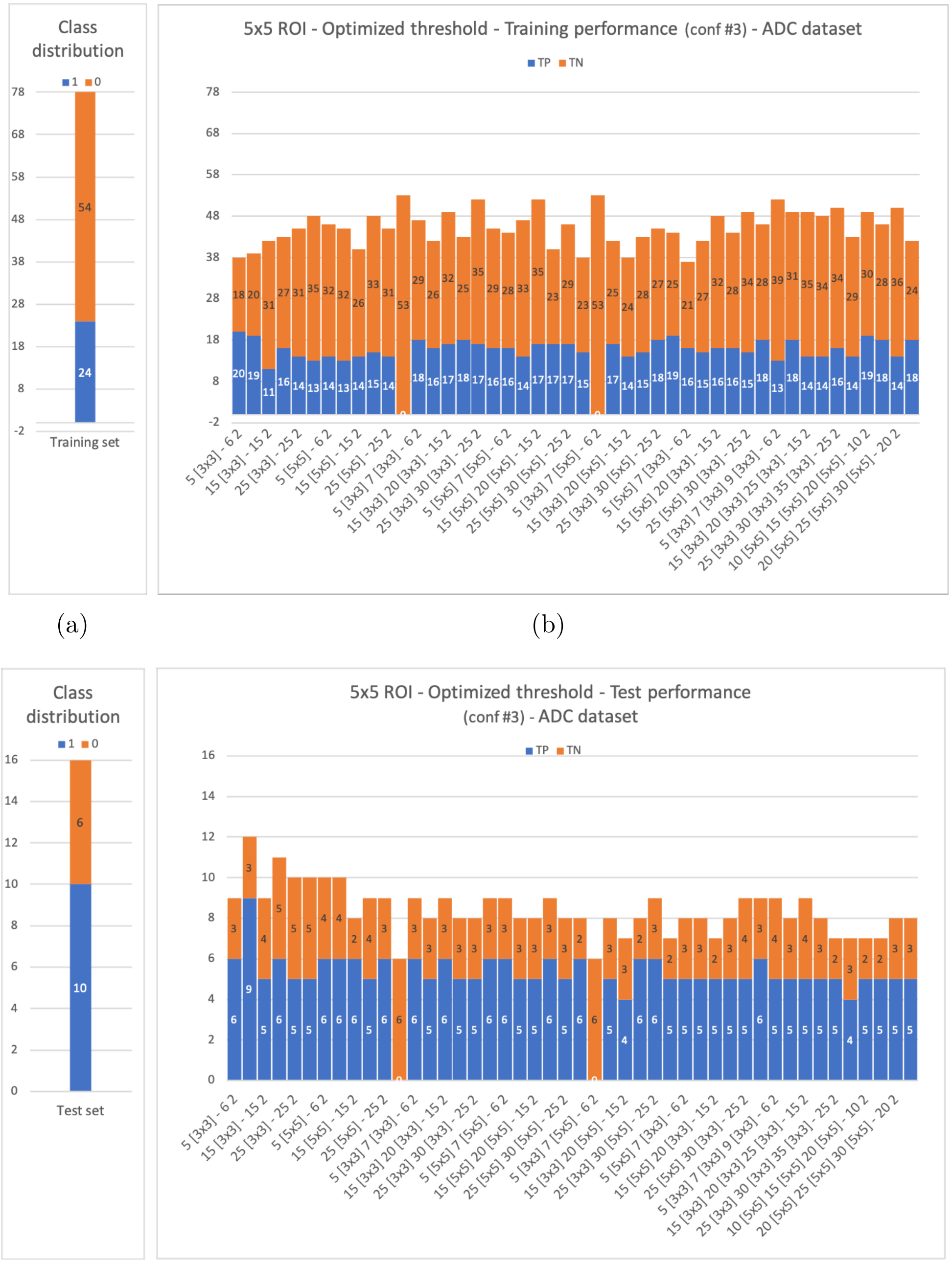
Figure 3.14: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, ADC dataset, and lesion prediction #2, i.e. with majority vote on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

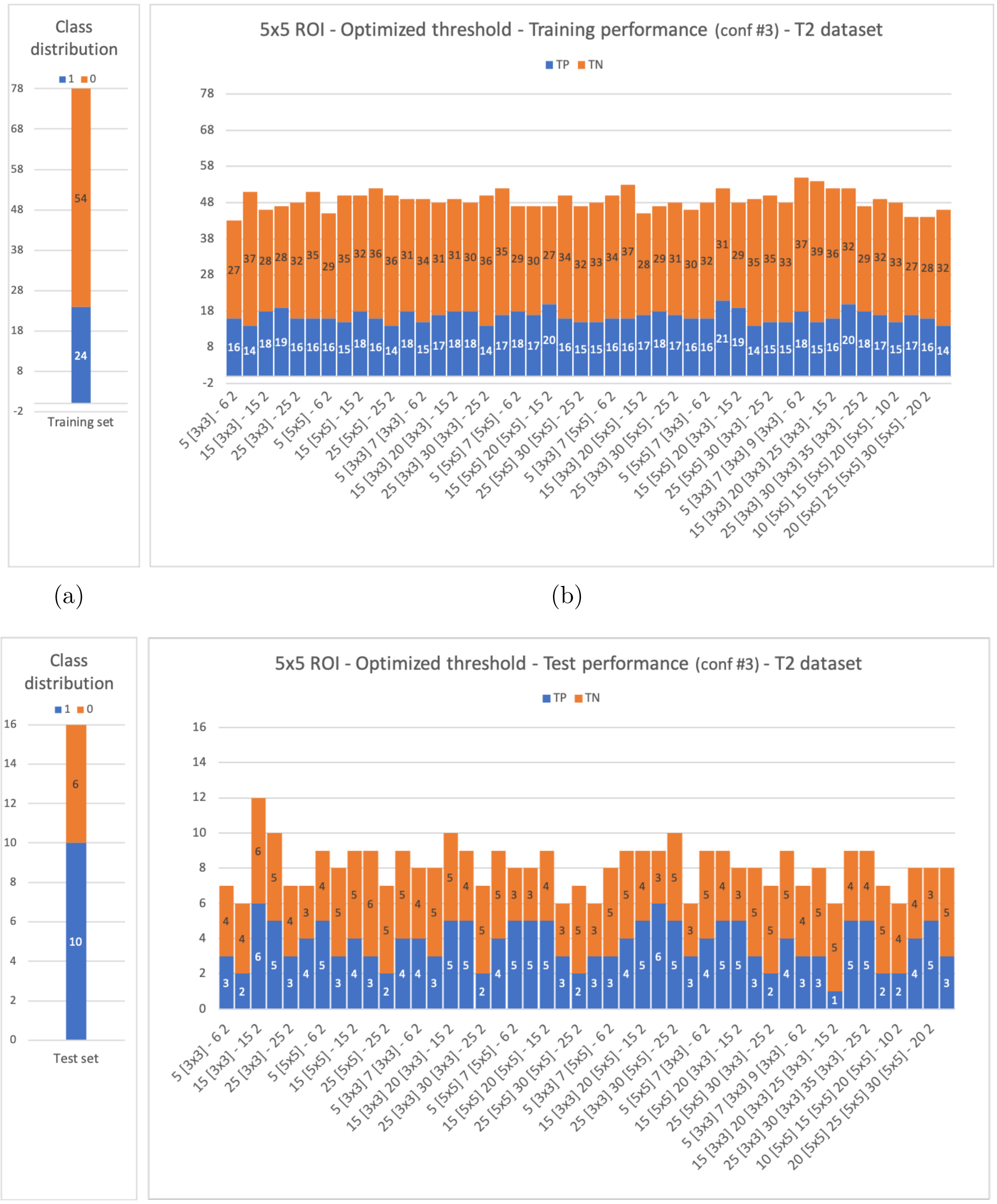
Figure 3.15: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, T2 dataset, and lesion prediction #2, i.e. with majority vote on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.16: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, ADC dataset, and lesion prediction #3, i.e. with optimed cut-off on ROIs and then on MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.17: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, T2 dataset, and lesion prediction #3, i.e. with optimed cut-off on ROIs and then on the MRI slices (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.

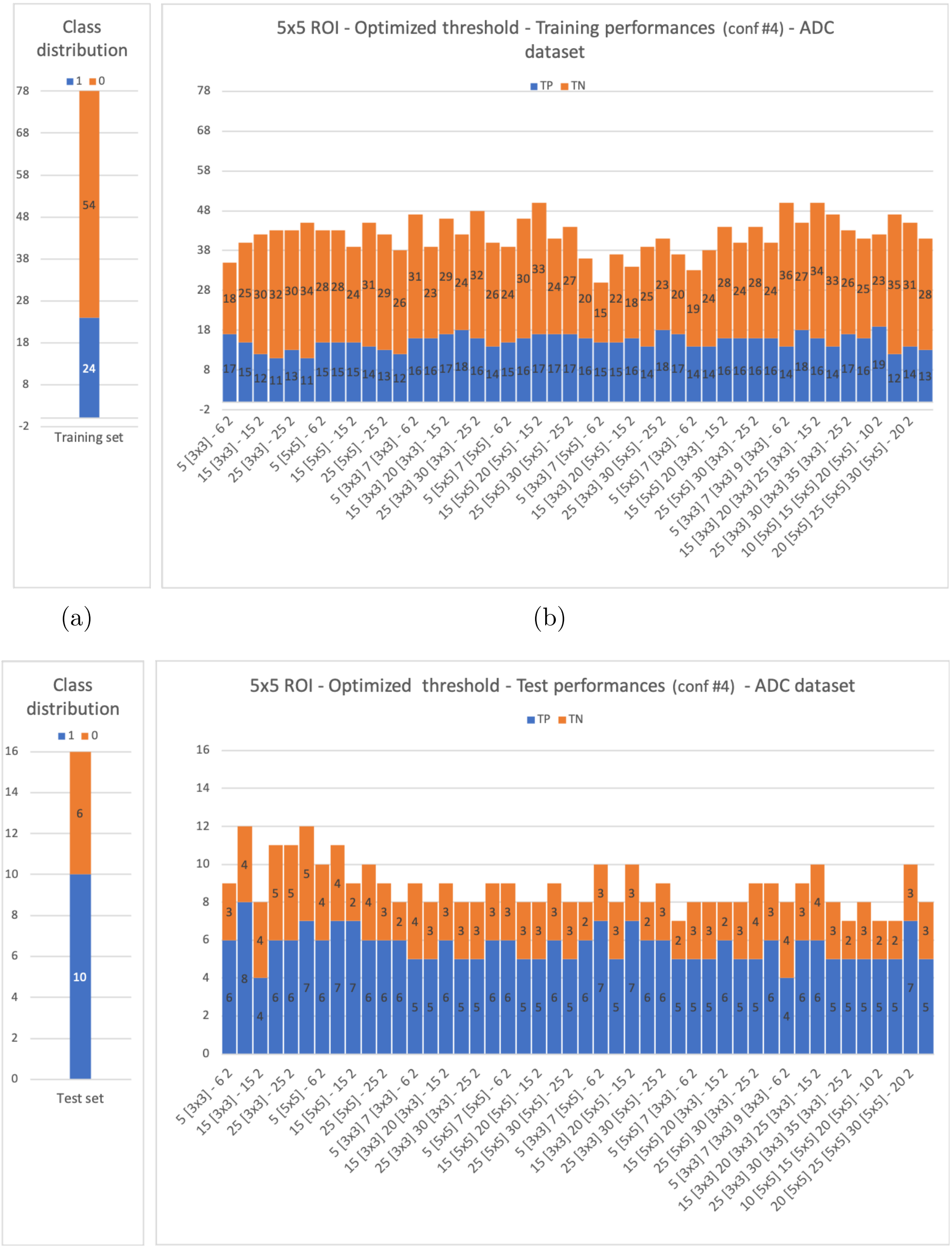
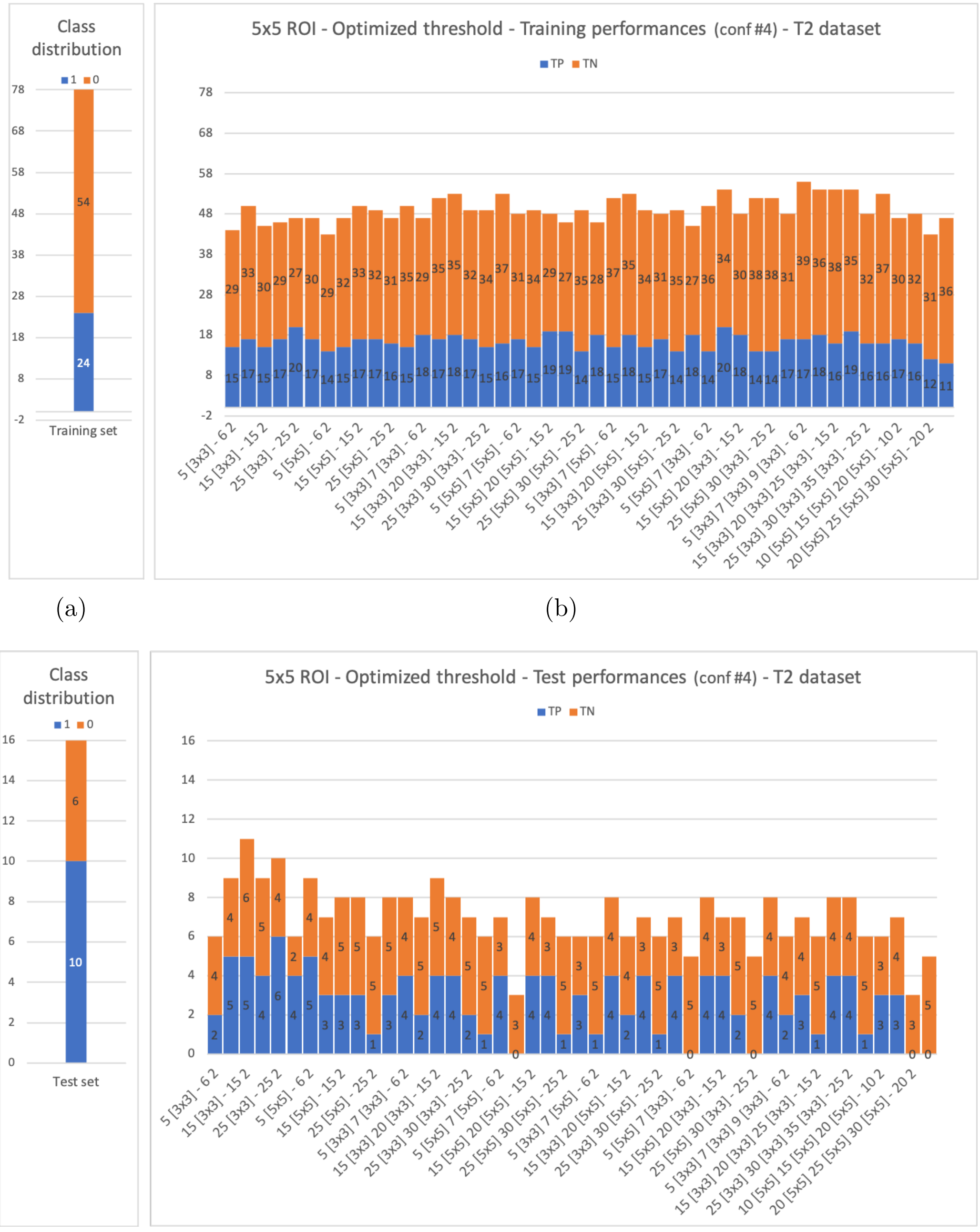


Figure 3.18: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, ADC dataset, and lesion prediction #4, i.e. with optimized threshold on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.



(a) (b)

(c) (d)

Figure 3.19: CNN results with training (b) and test (d) set, in terms of True Positive (TP) and True Negative (TN), for 5x5 ROI, T2 dataset, and lesion prediction #4, i.e. with optimized threshold on ROIs and biggest MRI slice prediction (see figure 3.7). (a) and (c) show the distribution of the two classes in training and test, respectively.

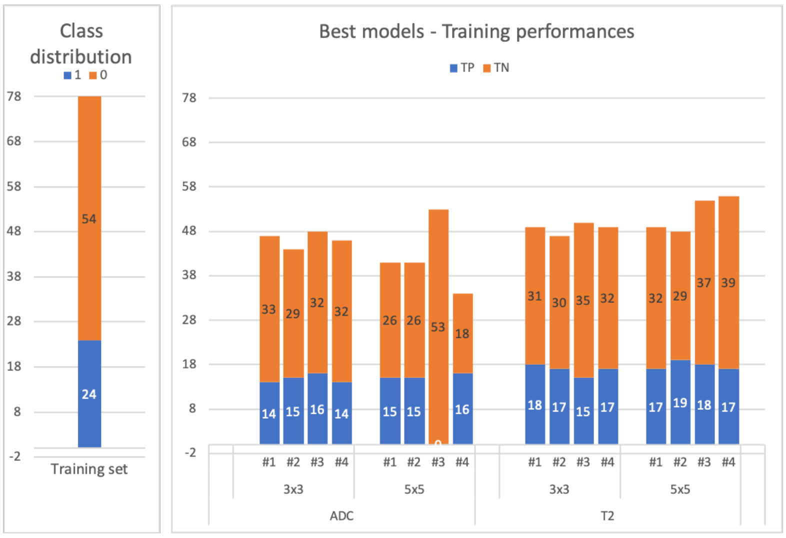
3.4.2 Best CNN structures: Validation set

By analyzing the results obtained on the training set, the models with the best performances are chosen. If more than one model is tied, only the one that performs best in the test set is considered. The structures corresponding to these best models are described in table 3.7.

Dataset	Prediction	ROI size (pixel)	
		3x3	5x5
ADC	#1	25 [3x3] 30 [3x3] - 25 2	5 [5x5] - 6 2
	#2	25 [3x3] 30 [3x3] - 25 2	5 [5x5] - 6 2
	#3	25 [3x3] 30 [3x3] - 25 2	30 [5x5] - 25 2
	#4	25 [3x3] - 25 2	15 [3x3] 20 [5x5] - 15 2
T2	#1	10 [3x3] 15 [3x3] - 10 2	10 [3x3] 15 [3x3] 20 [3x3] - 10 2
	#2	5 [3x3] - 6 2	10 [3x3] 15 [3x3] 20 [3x3] - 10 2
	#3	25 [3x3] 30 [3x3] - 25 2	5 [3x3] 7 [3x3] 9 [3x3] - 6 2
	#4	20 [3x3] - 20 2	5 [3x3] 7 [3x3] 9 [3x3] - 6 2

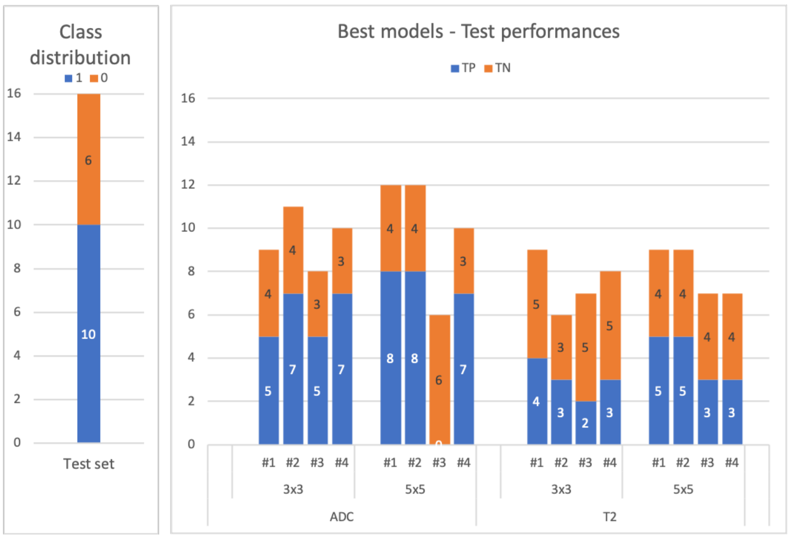
Table 3.7: Structures of the best Deep learning models in the four predictions described in the flowchart in figure 3.7, for ADC and T2 dataset. Number of filters [filter size] for each Convolutional layer, and number of neurons per Fully Connected Layer.

The performances of the best DL models are shown in figure 3.20, 3.21, and 3.22 in terms of TP and TN in training, test, and validation set, respectively. Observing the performances of the best models, it is interesting to note how the best two training models (T2, 5x5, # 3 and # 4), have low performances in the test set, but the highest in the validation. A similar situation to that observed in the chapter on machine learning (section 2.6). In general, however, T2 models have medium-low results in the test set, and tend to classify all lesions in the validation set as high-aggressive, thus obtaining high sensitivity but very low specificity. Conversely, ADC models perform better in the test set than in the validation set.



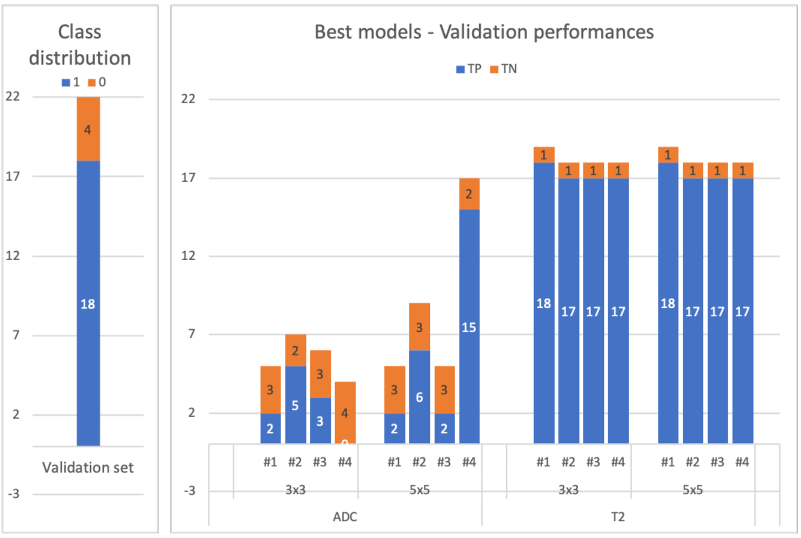
(a) (b)

Figure 3.20: Performance of the best DL models with the training set, in terms of True Positive (TP) and True Negative (TN), for each of the four lesion prediction methods (see figure 3.7).



(a) (b)

Figure 3.21: Performance of the best DL models with the test set, in terms of True Positive (TP) and True Negative (TN), for each of the four lesion prediction methods (see figure 3.7).



(a) (b)

Figure 3.22: Performance of the best DL models with the validation set, in terms of True Positive (TP) and True Negative (TN), for each of the four lesion prediction methods (see figure 3.7).

3.5 Discussion

The approach to CNN started by extracting the 3x3 and 5x5 pixel ROIs from each image, selecting only those windows completely inside the lesion, from all the MRI slices in which the tumor was present. Numerous CNN structures were tested, and different values for the parameters evaluated. The structures and parameters written here were found to be those that allowed the best performance of the models and that avoided overfitting.

The output of the CNNs provided the prediction of each ROI. Then, to obtain that of the MRI slice, a threshold was applied on the number of ROIs classified as high-aggressives. This was repeated for two thresholds: majority vote and optimized threshold (cut-off which allowed to obtain the best values of sensitivity and specificity). Once the slice predictions were obtained, the goal was to find the best way to aggregate these predictions at the lesion level. Most of the tumors had a section in several MRI slices, so it was decided to apply the same two thresholds applied at the ROI level. However, observing the MRI slices, some tumors had a very large section in one slice and a smaller one in the others. In these cases, applying a threshold on slice prediction put the prediction of larger and smaller tumor sections on the same level. Therefore, it was decided to predict the lesion class not only with a threshold applied to the slices but also by taking only the

prediction of the slice containing the largest tumor section.
The results obtained are not optimal. As in the case of machine learning, classifiers that perform best on the test set are the worst in the validation set. Once again, the presence of lesions coming only from Candiolo in the test set, and only from Molinette in the validation set should be considered.

Chapter 4

Machine and Deep Learning model comparison

This chapter shows the best models described in the two previous sections, regarding Machine and Deep learning. Table 4.1 shows, for machine learning (a), the performance of the best model of each FS method implemented, and, for deep learning (b), the best model performance for each of the four lesion prediction type strategies.

FS method		training			test			validation	
	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity
AP	1,00	1,00	1,00	0,75	0,70	0,83	0,64	0,56	1,00
AUC	1,00	1,00	1,00	0,69	0,70	0,67	0,68	0,67	0,75
GA	1,00	1,00	1,00	0,75	0,70	0,83	0,59	0,56	0,75
MRMR	0,97	0,96	0,98	0,50	0,50	0,50	0,82	0,83	0,75

(a)

Lesion prediction		training			test			validation	
	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity	Accuracy	sensitivity	specificity
#1	0,63	0,75	0,57	0,56	0,40	0,83	0,86	1,00	0,25
#2	0,63	0,75	0,57	0,56	0,50	0,66	0,82	0,94	0,25
#3	0,70	0,75	0,68	0,44	0,30	0,67	0,82	0,94	0,25
#4	0,72	0,71	0,72	0,44	0,30	0,67	0,82	0,94	0,25

(b)

Table 4.1: Performances of the best model obtained for the Machine Learning (a) and Deep learning (b) approach.

Interestingly, three out of four ML models (AP, AUC, GA) have been trained with 3D features and a Gaussian kernel (see table 4.2-(a)). Regarding the DL, all four CNN models come from the T2 dataset, specifically ROI 3x3 in the case of lesion prediction #1 and ROI 5x5 in the other three types of predictions (see table 4.2-(b)).

FS method	Dataset	kernel
AP	3D T2	polynomial
AUC	3D ADC-T2 joined	polynomial
GA	3D T2	polynomial
MRMR	2D ADC	Gaussian

(a)

Lesion prediction	Dataset	ROI size	CNN structure
#1	T2	[3x3] pixels	10 [3x3] 15 [3x3] - 10 2
#2	T2	[5x5] pixels	10 [3x3] 15 [3x3] 20 [3x3] - 10 2
#3	T2	[5x5] pixels	5 [3x3] 7 [3x3] 9 [3x3] - 6 2
#4	T2	[5x5] pixels	5 [3x3] 7 [3x3] 9 [3x3] - 6 2

(b)

Table 4.2: Structure specifications of the best models of Machine Learning (a) and Deep Learning (b).

Considering all the models shown in table 4.2, six out of eight come from T2W images, one from the ADC-T2 joined dataset, and only one comes from the ADC dataset. Therefore, it seems that models trained with T2W images are able to better differentiate the two classes.
In conclusion, machine learning models perform better in training and test sets than deep learning models, but the latter obtain higher performances in the validation set. This is probably due to the strong imbalance of the validation set. Indeed, DL models tend to classify new lesions in the high-aggressive class, and since 82 % of the validation set is composed of high-aggressive lesions, the resulting performance is quite high. In fact, this results in high sensitivity but very low specificity: such models seem not to be able to recognize low-aggressive tumors (only one out of four is correctly classified).

Chapter 5

Conclusion and Future Perspectives

The aim of this study was to characterize PCa aggressiveness through bi-parametric MRI images. Specifically, by carrying out an in-depth search of the literature about the classification of PCa Gleason Groups, it emerged that most of the researchers focused on distinguishing between clinically significant tumors (GS$>$ 6) and not (GS $<=$ 6). However, given the heterogeneity of PCa with GS 3+4 and 4+3, and the possibility of not over-treating and monitoring 3+4 tumors with active surveillance, it was considered useful to investigate the possibility of creating a CADx capable of classifying between low-aggressive and high-aggressive PCas. The use of bi-parametric MRI was preferred to multi-parametric as it allows to obtain images of the lesion in less time and with less invasiveness for the patient.

The approach was dual: first with machine learning and later with deep learning. To fully understand the results that have been obtained, it is necessary to make a premise. The dataset used is made up of lesions coming from two health facilities: Candiolo IRCCS and San Giovanni Molinette hospital. To proceed with the construction of the classifiers, the patients of the two hospitals were divided into training, test, and validation set. While the division of Candiolo's patients, in training and test sets, was carried out according to the size of the lesions, Molinette ones were divided according to temporal availability. In fact, Molinette patients were included in the study only later, in two steps: in the first, twenty-one patients were available, who were all included in the training set; in the second step a further forty-eight patients were obtained, which were all used to form an external validation set. This did not allow an optimal division into training, test and validation set, and is therefore one of the main limitations of the study. The models obtained, both in the ML and DL cases, did not manage to obtain high performances both in the test set and in the validation set. This is due to the fact that the test set is made up of Candiolo lesions only, while the valdation of patients from Molinette only.

Despite these limitations, the models achieve promising results. With regards to machine learning, the best model achieves excellent performance in the training set (100% accuracy), good results in the test set (75%, 70% and 83%, respectively of accuracy, sensitivity, and specificity) and results slightly lower in the validation set (64%, 56%, 100% of accuracy, sensitivity, and specificity respectively). Regarding the deep learning approach, the best model achieves good performance in the training set (71%, 72%, and 71% respectively of accuracy, sensitivity, and specificity), low performance in the test set (44%, 30%, and 67%, respectively of accuracy, sensitivity, and specificity) and slightly higher results in the validation set (82%, 94%, and 25% respectively of accuracy, sensitivity, and specificity).

In the future, it would be necessary to reconsider the division of patients into training, testing, and validation sets, in order to obtain a training set that is more representative of the variability of the two tumor classes. Furthermore, it would be interesting to evaluate additional Machine Learning models, as well as other configurations for CNN structures.

Acknowledgements

I would like to start by thanking my supervisor, Samanta Rosati, who gave me the opportunity to work on this project. A special thanks goes to my co-supervisor, Valentina Giannini: from the very first day, circumstances forced us to communicate remotely, and despite the resulting difficulties, she has always been ready to help me, advising me and following me during every phase of the project. Above all, I am grateful to her for believing in me and in my potential.

A special thanks goes to my parents: the gratitude is as big as the love I feel for them. They have been a reference, in every moment, even thousands of kilometers away. I thank my mother for teaching me that, if the road is long, it should not be looked at in its entirety, but must be faced one step at a time, and that it is right to know how to allow yourself to take a break to breathe. To my dad, who, with wise silence, has always known how to listen me and advise me. He is always the first to remind me of my worth. They are my pillars.

To my sister. My example from my birth. As a child I tried to imitate her. As we grew up, our personalities stood out, and, from imitating her, I moved on to take inspiration from her, from her strength, from her tenacity. I could not imagine a life without her.

To my grandmother, who, with her caresses on the face and her delicacious plates, has always protected me, making me away, even for a moment, from the frenzy that university life can lead to.

To Chiara, my longtime friend: no one can know me better than her. The distance tested us, but we won. Thank you for always supporting me in anytime, even with a simple phone call.

To Mihaela, the girl I met on the floors of the Polytechnic, on the first day of university. It helped me take the first steps towards a new life, alone, in a new city. I will never thank her enough for always being by my side.

To Thomas, the one who cheered me up and made me relax in post-study moments, always ready to listen to me (and accompanied by an excellent beer). I thank him for being a reference, a precious friend.
To Giulia. One of the people that I think is most similar to me. One of the friends I know I can always count on. With her I share not only the name, but also opinions, thoughts, and values. Thank you for always believing in me.
To Laura. She always made me feel like I was her daughter. I admire her, for her elegance and her sweetness, and I thank her, for the familiar warmth she has always made me feel.
To Miriana. Companion par excellence of my university path. It is with her that I have lived most of the pre-exam moment, and the post-celebrations. Thank you for helping me face everything with more lightness and more awareness, and for always being a present and honest friend.
To Roberta, whom I admire for her dedication to work. Thank you for the advice you have been able to give me, and for the support you have always shown me.

Thanks so much to all of you.
Yours,
Giulia.

List of Tables

List of Figures

Bibliography

[1] *What Is Prostate Cancer?* 2020. URL: https://www.cdc.gov/cancer/prostate/basic_info/what-is-prostate-cancer.htm (cit. on p. 2).

[2] O. Singh; S. R. Bolla. «Anatomy, Abdomen and Pelvis, Prostate». In: *StatPearls [Internet]* (Jan. 2020). URL: https://www.ncbi.nlm.nih.gov/books/NBK540987/ (cit. on pp. 1, 2).

[3] J. E. McNeal. «Anatomy of the prostate and morphogenesis of BPH». In: *Progress in clinical and biological research* (1984) (cit. on p. 1).

[4] Urology Match. *Basic Principles: Prostate Anatomy.* Accessed: 31-08-2020. URL: http://www.urologymatch.com (cit. on pp. 1, 2).

[5] N. Mottet et al. «Effect of patient age on early detection of prostate cancer with serum prostate-specific antigen and digital rectal examination». In: *Urology. 1993;42(4):365-374* (1993). DOI: 10.1016/0090-4295(93)90359-i (cit. on p. 1).

[6] European Association Urology. *European Association of Urology Pocket Guidelines. 2020 Edition.* Vol. presented at the EAU Annual Congress Amsterdam 2020. European Association of Urology Guidelines Office, 2020. ISBN: 978-94-92671-11-0. URL: http://uroweb.org/guidelines/compilations-of-all-guidelines/ (cit. on pp. 1, 4–7).

[7] R. B.Shah and M. Zhou. «Needle Biopsy Sampling Techniques and Role of Multiparametric-Magnetic Resonance Imaging Modality in Prostate Cancer Diagnosis and Management.» In: *Springer International Publishing* (2019). DOI: 10.1007/978-3-030-13601-7_2 (cit. on p. 3).

[8] J. Ferlay et al. «Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods». In: *Int. J. Cancer: 144, 1941–1953* (2019). DOI: 10.1002/IJC.31937 (cit. on p. 2).

[9] K. D. Miller. R. L. Siegel. and A. Jemal. «Cancer Statistics, 2020». In: *CA A Cancer J Clin, 70: 7-30* (2020). DOI: 10.3322/caac.21590 (cit. on p. 2).

[10] *American Cancer Society.* Accessed: 5-09-2020. URL: https://www.cancer.org/cancer/prostate-cancer (cit. on p. 2).

[11] G. Carioli et al. «European cancer mortality predictions for the year 2020 with a focus on prostate cancer». In: *Annals of Oncology* (2020). DOI: 10.1016/j.annonc.2020.02.009 (cit. on pp. 2, 4–6).

[12] T. Y. Chan et al. «Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy.» In: *Urology* (2000). DOI: 10.1016/s0090-4295(00)00753-6 (cit. on p. 8).

[13] J. L. Wright et al. «Prostate cancer specific mortality and Gleason 7 disease differences in prostate cancer outcomes between cases with Gleason 4 + 3 and Gleason 3 + 4 tumors in a population based cohort.» In: *Journal of Urology* (2009). DOI: 10.1016/j.juro.2009.08.026 (cit. on p. 8).

[14] J. R. Stark et al. «Gleason Score and Lethal Prostate Cancer: Does 3 + 4 = 4 + 3?» In: *Journal of clinical oncology* (2009). DOI: 10.1200/JCO.2008.20.4669 (cit. on p. 8).

[15] J. I. Epstein et al. «A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score». In: *European Urology* (2016). DOI: 10.1016/j.eururo.2015.06.046 (cit. on p. 8).

[16] C. J. Kane et al. «Variability in Outcomes for Patients with Intermediate-risk Prostate Cancer (Gleason Score 7, International Society of Urological Pathology Gleason Group 2-3) and Implications for Risk Stratification: A Systematic Review». In: *European Urology Focus* (2017). DOI: 10.1016/j.euf.2016.10.010 (cit. on p. 8).

[17] J. I. Epstein et al. «Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System». In: *Am J Surg Pathol. 40(2):244-252* (2016). DOI: 10.1097/PAS.0000000000000530 (cit. on p. 8).

[18] Jungheum Cho et al. «Biparametric versus multiparametric magnetic resonance imaging of the prostate: detection of clinically significant cancer in a perfect match group». In: *Prostate International* (2020). DOI: 10.1016/j.prnil.2019.12.004 (cit. on p. 9).

[19] M. D. Greer et al. «Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study». In: *European Radiology* (2018). DOI: 10.1007/s00330-018-5374-6 (cit. on p. 9).

[20] T. Hambrock et al. «Prostate cancer: computer-aided diagnosis with multiparametric 3-T MR imaging–effect on observer performance». In: *Radiology* (2013). DOI: 10.1148/radiol.12111634 (cit. on p. 9).

[21] V. Giannini et al. «Multiparametric magnetic resonance imaging of the prostate with computer- aided detection: experienced observer performance study.» In: *European Radiology* (2017). DOI: `10.1007/s00330-017-4805-0` (cit. on p. 9).

[22] G. Lemaître et al. «Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review». In: *Computers in Biology and Medicine* (2015). DOI: `10.1016/j.compbiomed.2015.02.009` (cit. on pp. 9, 11).

[23] R. R. Wildeboer et al. «Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods.» In: *Computer Methods and Programs in Biomedicine* (2020). DOI: `10.1016/j.cmpb.2020.105316` (cit. on p. 9).

[24] R. Cao et al. «Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet». In: *IEEE Trans Med Imaging* (2019). DOI: `10.1109/TMI.2019.2901928` (cit. on p. 10).

[25] D. Fehr et al. «Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images». In: *National Academy of Sciences* (2015). DOI: `10.1073/pnas.1505935112` (cit. on p. 10).

[26] P. Tiwari et al. «Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS». In: *Medical Image Analysis* (2012). DOI: `10.1016/j.media.2012.10.004` (cit. on p. 10).

[27] C. Fusun et al. «Final Gleason Score Prediction Using Discriminant Analysis and Support Vector Machine Based on Preoperative Multiparametric MR Imaging of Prostate Cancer at 3T.» In: *BioMed research international* (2014). DOI: `10.1155/2014/690787` (cit. on p. 10).

[28] A. Vignati et al. «Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness.» In: *Physics in Medicine Biology* (2015). DOI: `10.1088/0031-9155/60/7/2685` (cit. on p. 12).

[29] G. Nketiah et al. «T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results.» In: *European Society of Radiology* (2016). DOI: `10.1007/s00330-016-4663-1` (cit. on p. 12).

[30] T. W. Baek et al. «Texture analysis on bi-parametric MRI for evaluation of aggressiveness in patients with prostate cancer.» In: *Abdominal Radiology* (2020). DOI: `10.1007/s00261-020-02683-4` (cit. on p. 12).

[31] C. Jensen et al. «Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier.» In: *Medical Imaging* (2019). DOI: `10.1002/acm2.12542` (cit. on p. 12).

[32] M. J. Kucharczyk A. Chaddad and T. Niazi. «Multimodal Radiomic Features for the Predicting Gleason Score of Prostate Cancer.» In: *Cancers* (2018). DOI: `10.3390/cancers10080249` (cit. on p. 12).

[33] S. Chakraborti J.D. Gibbson. *Nonparametric Statistical Inference, Fifth Edition.* IBN-13: 978-1-4200-7762-9. 2011 (cit. on p. 17).

[34] J. Lee et al. «Deep Learning in Medical Imaging: General Overview». In: *Korean Journal of Radiology* (2017). DOI: `10.3348/kjr.2017.18.4.570` (cit. on pp. 59, 61).

[35] R. Yamashita et al. «Convolutional neural networks: an overview and application in radiology». In: *Insights Imaging* (2018). DOI: `10.1007/s13244-018-0639-9` (cit. on p. 61).