BIOMEDICAL ENGINEERING MSc PROGRAM
UNIVERSITY OF CRETE
TECHNICAL UNIVERSITY OF CRETE
FOUNDATION FOR RESEARCH AND TECHNOLOGY-HELLAS

Application Grade Thesis

Title: Developing a Comprehensive Radiomic Analysis Workflow for the Detection of Prostate Cancer Aggressiveness on T2weighted MR Prostate Data

Τίτλος: Ανάπτυξη μοντέλων ραδιομικής ανάλυσης για την ανίχνευση της επιθετικότητας του καρκίνου του προστάτη σε T2w εικόνες προστάτη μαγνητικής τομογραφίας

Student's Name: Stylianos Zafeiris

Supervisor's Name: Konstantinos Marias

Co-supervisors' Names: Michalis Zervakis, George Manikis

Date of completion: 26 June 2023

This dissertation is submitted as a partial fulfilment of the requirements for the Master's degree of Biomedical Engineering M.Sc. program

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Kostas Marias and co-supervisor Prof. Michalis Zervakis for his helpful advice, relentless support and guidance throughout my thesis. I am also extremely grateful to Dr. George Manikis for his co-supervision. His unwavering guidance, invaluable suggestions and extensive knowledge assisted me throughout my thesis.

This endeavour would not have been possible without the helpful contributions of Katerina Dovrou. Her practical suggestions, knowledge and constructive criticism were paramount for completing my thesis. I would like to extend my sincere thanks to all my close friends for always encouraging me and being here for me.

Finally, I could not have undertaken this journey without the constant and unconditional support of my family. Words cannot express my gratitude to my parents George and Ria for their profound belief in my abilities and patience that cannot be underestimated.

To my grandparents

Abstract

Prostate cancer (PCa) is the second most common cancer diagnosed in male population worldwide, affecting 1.4 million men annually. Early assessment of the malignancy is crucial for treatment planning and extending patients' life expectancy. Imaging modalities such as Magnetic Resonance Imaging (MRI) are used for the non-invasive classification of patients in order to prevent overtreating indolent malignancies and undertreating those who warrant immediate treatment. The field of radiomics offers a large quantity of imaging features that describe the cancer phenotype and can be used in training machine learning (ML) models for predicting cancer aggressiveness. Effective model training necessitates feature selection, decreasing the high dimensionality and ensuring the inclusion of pertinent and non-redundant features. The objective of this study is to investigate the most commonly used feature selection methods and classifiers in order to predict the tumor's aggressiveness and analyze how various image preprocessing techniques affect the performance of the models. A publicly available multivendor dataset consisting of 225 samples with clinically significant PCa (csPCa) from 220 patients was used for the analysis. Samples were split in two cohorts based on ISUP score provided by clinicians. The first cohort (*n = 135*) contains samples with an assigned ISUP score equal to 2 (low aggressiveness csPCa) and the second cohort (*n = 90*) comprise samples with an assigned ISUP score of 3, 4 and 5 (high aggressiveness csPCa). Samples with ISUP score equal to 2 tend to have cancer cells that grow slowly, as opposed to the moderate and quick growth of cancer cells in samples of the second cohort. Thus, early detection of the tumor grade could prevent an unnecessary intervention or accelerate biopsy. A comprehensive search for the optimal pipeline was conducted for classifying the aggressiveness of csPCa. Intensity normalization methods and the N4 bias field correction method were used to investigate whether these preprocessing steps affect the performance of the models. For the original and each pre-processed dataset, a cross-combination strategy leveraging 6 classifiers and 13 feature selection methods was used for determining an optimal pipeline that reduces overfitting and best determines the tumor grade. Furthermore, hybrid feature selection methods were also investigated, using the optimal parameter set extracted from the pipeline. Methods investigated in this study demonstrated a balanced accuracy of 70% in determining the tumor's aggressiveness, providing promising results in early detection of aggressiveness of csPCa.

Περίληψη

Ο καρκίνος του προστάτη είναι ο δεύτερος πιο συχνός καρκίνος που διαγιγνώσκεται στον ανδρικό πληθυσμό παγκοσμίως, επηρεάζοντας 1,4 εκατομμύρια άνδρες ετησίως. Η πρώιμη αξιολόγηση της κακοήθειας είναι κρίσιμη για τον σχεδιασμό της θεραπείας και την επέκταση του προσδόκιμου ζωής των ασθενών. Οι απεικονιστικές μέθοδοι, όπως η Μαγνητική Τομογραφία, χρησιμοποιούνται για την μη-επεμβατική κατηγοριοποίηση των ασθενών, ώστε να αποφασιστεί η κατάλληλη θεραπεία τους ανάλογα με τον βαθμό κακοήθειας. Ο τομέας της ραδιομικής προσφέρει μία μεγάλη ποσότητα απεικονιστικών χαρακτηριστικών που περιγράφουν το φαινότυπο του καρκίνου και μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων μηχανικής μάθησης, με σκοπό την πρόβλεψη του βαθμού της επιθετικότητας του όγκου. Η αποδοτική εκπαίδευση τέτοιων μοντέλων προϋποθέτει την ύπαρξη μεθόδων επιλογής χαρακτηριστικών, ώστε να μειωθεί η υψηλή διάσταση του χώρου των χαρακτηριστικών και να διασφαλιστεί η επιλογή συναφών και χρήσιμων χαρακτηριστικών για την πρόβλεψη. Στόχος της παρούσας μελέτης είναι η χρήση και η αξιολόγηση των πιο ευρέως χρησιμοποιούμενων μεθόδων επιλογής χαρακτηριστικών και ταξινομητών, προκειμένου να προβλεφθεί η επιθετικότητα του καρκίνου, καθώς και η ανάλυση της επιρροής διάφορων μεθόδων προ-επεξεργασίας εικόνων στην απόδοση των μοντέλων. Ένα δημόσια διαθέσιμο σύνολο δεδομένων με εικόνες από διαφορετικούς προμηθευτές μαγνητικών τομογράφων, το οποίο περιλαμβάνει 225 δείγματα με κλινικά σημαντικό καρκίνο του προστάτη από 220 ασθενείς, χρησιμοποιήθηκε για την ανάλυση. Τα δείγματα διαχωρίστηκαν σε δύο ομάδες βάση του σκορ ISUP το οποίο εξήγαγαν κλινικοί. Η πρώτη ομάδα δειγμάτων (*v=135*) συμπεριλαμβάνει τα δείγματα με ISUP σκορ ίσο με 2 (λιγότερο επιθετικός, αλλά κλινικά σημαντικός καρκίνος του προστάτη), ενώ η δεύτερη ομάδα (*v=90*) περιέχει τα δείγματα με ISUP σκορ ίσο με 3, 4, ή 5 (αρκετά επιθετικός, κλινικά σημαντικός καρκίνος του προστάτη). Ο διαχωρισμός αυτός έγινε με βάση το γεγονός ότι τα δείγματα της πρώτης ομάδας έχουν καρκινικά κύτταρα όπου αναπτύσσονται πιο αργά σε αντίθεση με την μεσαία προς ραγδαία ανάπτυξη των κυττάρων στα δείγματα της δεύτερης ομάδας. Συνεπώς, η έγκαιρη ανίχνευση του βαθμού της επιθετικότητας του καρκίνου μπορεί να αποτρέψει μια περιττή επέμβαση ή να επιταχύνει την βιοψία. Μια εμπεριστατωμένη αναζήτηση των βέλτιστων μεθόδων διεξήχθη, με σκοπό την κατηγοριοποίηση της επιθετικότητας του κλινικά σημαντικού όγκου. Μέθοδοι κανονικοποίησης των τιμών της έντασης της εικόνας καθώς και η μέθοδος N4 φιλτραρίσματος χρησιμοποιήθηκαν, για να εξεταστεί ο τρόπος επιρροής τους στην απόδοση των μοντέλων. Για το αρχικό σύνολο δεδομένων και τις επεξεργασμένες εκδοχές του χρησιμοποιήθηκαν συνδυασμοί από 6 ταξινομητές και 13 μεθόδους επιλογής χαρακτηριστικών, για να καθοριστούν οι βέλτιστες παράμετροι των μοντέλων, οι οποίες μειώνουν την πιθανότητα overfitting και ταυτόχρονα αυξάνουν την ικανότητα του μοντέλου να διαχωρίζει το βαθμό επιθετικότητας του καρκίνου του προστάτη. Επιπροσθέτως, χρησιμοποιήθηκαν υβριδικές μέθοδοι επιλογής χαρακτηριστικών, με βάση τις βέλτιστες παραμέτρους όπου εξήχθησαν από την αρχική ανάλυση. Οι μέθοδοι που εξετάστηκαν στην παρούσα έρευνα έδειξαν ισορροπημένη ακρίβεια (balanced accuracy) 70% για τον καθορισμό του βαθμού του όγκου, παρέχοντας ελπιδοφόρα αποτελέσματα για την πρώιμη ανίχνευση του επιθετικού, κλινικά σημαντικού καρκίνου του προστάτη.

Table of Contents

List of figures

List of tables

Chapter 1: Introduction

The second most common cancer diagnosed in male population is prostate cancer (PCa) which affects 1.4 million men worldwide annually [1], [2]. The prevalence of prostate cancer exhibits significant geographical variation with higher rates reported in developed countries, such as North America, Europe, and Australia, and lower rates in less developed countries. Disparities are related to the access to healthcare, socioeconomic factors, health illiteracy and genomic susceptibility. Well-established risk factors of the disease are age (PCa is diagnosed mostly in men over 50 years old), ethnicity, family history of the disease, and genetic mutations, i.e. genes strongly associated with PCa, such as BRACA1 and BRACA2 which influence the risk for PCa. In addition, modifiable factors, such as lifestyle, environment (e.g. exposure to chemicals and ionizing radiation) and alimentary factors can also influence the development of PCa. For instance, the intake of lycopene or soy reduce PCa risk [3], [4]. In contrast, physical inactivity, alcohol and dairy product consumption are associated with increased PCa risk [1], [5]. Tobacco is positively associated with deaths from PCa [1].

PCa originates in the prostate gland, which is located below the bladder and in front of the rectum. Prostate consists of three areas: a) the peripheral zone (PZ) constituting the majority of the prostate gland; b) the transition zone (TZ) encompassing the urethra; and c) the central zone (CZ). Most cases of PCa are found in PZ which is exactly behind the rectal wall and thus it is detectable with a rectal exam (RE). A common biomarker used in PCa screening is high prostate specific antigen (PSA) values in blood serum. PSA is a glycoprotein expressed by the prostate tissue, it can be found mostly in semen and it ordinarily circulates in the blood. This protein is produced by both cancerous and non-cancerous tissue and can be detected primarily using PSA tests. When PSA tests and RE indicate prostate cancer, a biopsy is performed to determine the aggressiveness of the tumor. The aggressiveness is measured using the Gleason score (GS) and the International Society of Urological Pathology (ISUP) grade group. However, investigating only PSA level is not enough for diagnosing PCa, since PSA tests demonstrate limited sensitivity and they cannot detect tumor's aggressiveness [6] neither distinguish between other prostate conditions i.e. prostatitis and prostatic hyperplasia [7]. The main cause of the limited diagnostic ability of PSA tests is that the levels of PSA in blood are influenced by external factors, such as lifestyle, hormonal profile, obesity and even infections in the urinary tract. Hence, these factors may lead to false positive or false negative results. Furthermore, similar PSA levels have been observed in patients with both low and high-risk tumors, making PSA a less accurate diagnostic tool. Even though screening using RE and PSA tests is recommended for early detection of cancer, other diagnostic tools could be used in the process of medical decision making for avoiding undertreatment of malignant tumors and overtreatment of indolent malignancies.

Imaging modalities are of paramount importance in diagnosing PCa. Magnetic Resonance Imaging (MRI) scans of the prostate is a ubiquitous screening method for the malignancy

detection, alongside with the RE and PSA tests. MRI results are interpreted by clinicians which can request immediate biopsy (if PI-RADS > 3), plan a treatment tailored to the patient's needs and monitor disease progress. In more aggressive forms, PCa manifests rapid growth with high possibility of metastasis to near organs, bones and lymph nodes. MRI demonstrates restricted capability of detecting metastasis, especially to lymph nodes. Thus, detection of the malignancy in early, less aggressive stage is crucial for the survival of the patient. The rapid development of the field of radiomics could aid early detection. Radiomic analysis entail high-throughput feature extraction (e.g., shape, texture, etc.), from medical images coupled with machine learning techniques for developing powerful diagnostic models. The non-invasive nature of the analysis and the promising results reported in the literature have led to an increased scientific interest in developing models for detection, segmentation, classification of the tumor and prediction of treatment response.

Developing radiomic analysis workflows is not an effortless task. Medical images require some preprocessing steps before being used in machine learning pipelines. The most common steps are anonymization for ensuring privacy of the patients, filtering and noise reduction for enhancing the quality of the images and feature extraction. Data normalization is a prerequisite for multicenter datasets due to the variability introduced to the images by different scanner vendors, models and acquisition protocols. In un-normalized datasets, the feature extraction process may derive imaging features of increased variability which may hamper robustness and generalizability of the machine learning models. Moreover, image filtering is crucial for attenuating variations in the pixel values of the images and thus enhancing drastically their quality. In magnetic resonance imaging, bias field correction is mandatory for correcting the intensity inhomogeneities, caused by the radiofrequency coil, eddy currents and several patient-related factors [8].

Another important step is feature selection, which reduces the number of features feeding only pertinent and non-redundant features to the machine learning (ML) models. It is a required step of the pipeline, since the high dimensionality of the feature space hinders the training process and reduces the predictive power of the models, a phenomenon known as the "Curse of Dimensionality". Feature selection methods can be divided in three main categories: filtering, embedded and wrapper [9]. Filtering methods can be further divided in univariate and multivariate methods. The former assumes that features are independent and selects features based only on the information they provide for the target class. The latter considers between-features interactions, selecting features that are highly correlated with the target class and provide the least information about the already selected features. Filter methods can evaluate the quality of the selected feature set quickly and are independent of the classifier used, which may cause reduction in the accuracy of the predictions. Wrapper feature selection methods select features iteratively, keeping features that maximize the classifier's performance. Even though these methods are slower, they demonstrate reduced classification error. Embedded methods (e.g. Random Forest features importance and Least Absolute Shrinkage and Selection Operator coefficients) are based on the intrinsic properties

of the classifier for calculating feature scores, resulting in higher execution speed. Furthermore, hybrid methods leveraging two or more feature selection methods are used in order to combine the strengths of each method acquiring an optimal feature set. Ensemble methods leverage multiple feature selection methods in parallel, improving the quality and stability of the selected features.

The current study investigates pipelines for predicting PCa's aggressiveness and analysing the influence of image preprocessing techniques on models' performance. A public dataset containing T2weighted images of the prostate from multiple vendors is used for the analysis. All images contain incidents of clinically significant PCa (csPCa) and are split in two cohorts based on cancer aggressiveness, indicated by the ISUP score. Most common feature selection methods and classifiers from the literature are evaluated in different combinations for determining an optimal pipeline and parameter set. Intensity normalization methods and the N4 bias field correction method are applied to the images for investigating their impact on the performance of the models. In addition, hybrid feature selection methods are investigated using the optimal parameter set extracted from the previous analysis. An unseen hold-out test set of the initial dataset is used for evaluating the models' performance.

A comprehensive description of the analysis is presented in the next chapters. Chapter 2 introduces an extensive literature review, concerning the recent breakthroughs in the field of image preprocessing, radiomics, feature selection and classification in prostate cancer. Chapter 3 describes the methodology followed in this study and chapter 4 reports the results of the analysis. A discussion about the findings of the study and a brief comparison with other studies is presented in chapter 5 and finally, chapter 6 concludes the study.

Chapter 2: State-of-the-art

The high prevalence of prostate cancer in men worldwide has drawn the attention of many researchers towards the detection of the tumor's aggressiveness. Prostate cancer aggressiveness classification is a critical task that assists clinicians in disease management through guiding treatment planning and possibly predicting patient outcomes. Accurate classification allows clinicians to differentiate between indolent tumors that may require conservative approaches, such as active surveillance, and malignancies that demand immediate action. Several biological methods are employed for diagnosing PCa aggressiveness. Assessment of histopathological features after the prostate tissue biopsy is the baseline approach. However, several molecular markers and genetic profiling techniques involving the analysis of specific gene mutations and expression are gaining prominence for assessing the disease aggressiveness.

The most common molecular marker used for the diagnosis of the disease is PSA tests which measure the PSA values in the serum. Another common marker is PSA density (PSAd), which is calculated as the ratio of PSA to the volume of the whole gland [10]–[12]. Other PSA derivatives are RSA velocity [13] and PSA doubling time [14] were proposed; however, these biomarkers could not provide valuable insight for the outcome prior to the biopsy [15]. For improving the low sensitivity and specificity of measuring PSA values in the serum, age specific reference ranges are introduced in clinical practice for counterbalancing the influence of age and prostate volume on PSA values [16]. Horoszewicz et al. in 1987 [17] identified that Prostate-specific Membrane Antigen (PSMA) glycoprotein found in epithelial cells and blood was overexpressed in patients with PCa. A nuclear structural protein called Early Prostate Cancer Antigen (EPCA) is associated with cancer and has demonstrated high sensitivity and specificity in prostate cancer detection [18], [19]. A comprehensive review of molecular markers has been conducted by Bradford et al. [20]. Finally, Choudhury et al. [21] reviewed the use of several genetic markers in prostate cancer diagnosis and treatment planning.

Imaging has also a crucial role in diagnosing and managing patients with prostate cancer. Magnetic resonance imaging (MRI) is widely used in clinical practice for the detection and grading of the tumor. Prostate Imaging and Reporting and Data System (PI-RADS) [22] is a structured category assessment system developed in 2012 (version 1), containing clinical guidelines on a consensus basis for evaluating prostate multi-parametric MRI (mpMRI). It was updated in 2014 (version 2) [23], [24] for alleviating the confusion on how to weight each parameter of the mpMRI and achieving optimal prostate lesion characterisation. More precisely, PI-RADS score shows the likelihood of a tumor to be clinically significant cancer. The value of this score is determined based on the findings of the multiparametric MRI and the range of values is from 1 to 5, where higher values indicate higher risk of clinically significant cancer to be present. However, the PI-RADS score measures the likelihood of a tumor to be malignant rather than the tumor's aggressiveness. To this end, the International Society of Urological Pathology (ISUP) grading system and Gleason score are used to grade the tumor,

indicating the aggressiveness [25]. The ISUP grading system defines the grade of tumor with a value from 1 to 5 depending on the Gleason score [26] in order to better predict the prostate cancer outcome.

The emergence of the field of radiomics influenced the development of machine learning and deep learning algorithms for the automatic segmentation of the prostate gland and the lesions, and for the classification of the tumor's grade. The lesion segmentations are required for extracting radiomic features and they highly affect the values of the extracted features. Radiomic features are imaging features that describe the shape, the intensities distribution and the intensities dependencies, reflecting the tumor's heterogeneity.

Image quality and data harmonization are also important factors for radiomic feature extraction. Several image preprocessing techniques are employed for enhancing image quality and aim to address poor image quality, artifacts and standardize data for subsequent analysis, such as classification and radiomic feature extraction. Common preprocessing methods are bias field correction, intensity normalisation, resampling, filtering and discretization.

MRI images often suffer from intensity variations caused by the non-uniformity of the radiofrequency (RF) field during image acquisition. This low frequency signal, called bias field, degrade the image quality resulting in intensity inhomogeneities. Bias field correction techniques, such as N4 or nonparametric methods, are used to correct the intensity variations and improve the uniformity of the image. The N4 method [8] is a popular bias field correction method and has been used as a preprocessing step in classification [27] and segmentation studies [28]–[32]. However, this method has several parameters that their values should be tuned. Martin et al. [32] experimented on the values of some N4 parameters by applying the algorithm to breast phantoms. They identified 50 iterations, fitting level 5 and the use of a full mask as optimal configuration for the bias field correction reducing the intensity inhomogeneities. A recent study by Dovrou et al. [33] investigated a variable set of values for five parameters of N4ITK filter for bias field correction in MR prostate images. They used the Full Width at Half Maximum (FWHM) of the periprostatic fat distribution as metric to quantify the improvement of the image quality after applying the bias field correction method. The main hypothesis was that after applying the bias field correction, the tissue representation becomes more homogeneous and thus the value of the FWHM of the periprostatic fat distribution is smaller. They examined 240 different configurations of N4 bias field correction in 4 datasets with images scanned by surface coil and a combination of endorectal and surface coil at 1.5T and 3T magnetic field strength. The derived optimal configuration of the N4 filter was affected by the type of the coil used during the scanning of the subject rather than the magnetic field strength. The optimal configuration for images scanned with a combined surface and endorectal coil at 1.5T or 3T is: convergence threshold 0.001, shrink factor 2, fitting level 6, number of iterations 100 and the use of default mask. The optimal configuration for prostate images scanned with surface coil at 1.5T or 3T is: convergence threshold 0.001, shrink factor 2, fitting level 5, number of iterations 25 and the use of default mask.

MR images do not have standardized intensity values for the various tissues, hampering the direct comparison of MR images even when they are scanned with the same conditions and acquisition protocols. Intensity normalization techniques are employed to standardize the intensity levels across different MRI scans or sequences. This ensures consistency in the image intensities, allowing for more accurate comparison and analysis of the data. In the medical image analysis, same tissues should have the same intensity representation in order to be comparable for subsequent analysis. Hence, intensity normalization techniques aim to harmonize the intensities of MR images and bring them into a common scale. The most common intensity normalization techniques used in MR images are min-max scaling, z-score normalization and the histogram matching method [34]. Nyul et al. [34] introduced a non-linear histogram normalization technique for image harmonization. This technique learns a standard histogram from a set of images, identifying specific landmarks and then linearly maps the image intensities to the intensities of the standard histogram. The landmarks are histogram-specific parameters that describe the distribution of the histogram.

Furthermore, resampling involves changing the resolution or voxel size of the MRI image, which can be useful for matching the resolution of different images or facilitating computational analysis. Discretization is the process where the signal intensities are clustered to specific range intervals, i.e. bins of the histogram, in order to limit the range of intensities. This is a crucial step to efficiently calculate the radiomic features, reducing the computational complexity. There are two major categories of discretization: a) the absolute discretization with fixed bin size/width (FBS) method and the relative discretization with fixed bin number (FBN) method. There is no agreement in which method is better and it may be application specific.

Feature selection techniques are used for reducing the computational complexity and avoiding overfitting of the models. Quantitative features extracted from images are widely used in several machine learning approaches. The dimensionality of the feature space is high, requiring increased computational complexity for processing and model training. A common category of feature selection techniques is univariate filtering methods, such as Pearson and Spearman correlation-based feature selection. MRMR [35] is a multivariate filtering method, commonly used in the literature. Embeded methods like Least Absolute Shrinkage and Selection Operator (LASSO) [36], wrapper methods like Boruta [37] and the Relief family of feature selection algorithms [38] are widely used as feature selection methods.

Several studies have investigated the ability of radiomic features to predict the prostate cancer aggressiveness. Chaddad et al. [39] investigated the ability of radiomic features extracted from T2W images and Apparent Diffusion Coefficient (ADC) maps to non-invasively predict the Gleason score. The workflow of the analysis included the detection of the sub-volume of the Region of Interest (ROI), the radiomics extraction and the statistical and machine learning analysis and is presented in Figure 1. In this study, 99 patients with prostate cancer were included and 41 radiomic features were extracted from the tumor sub-volume. The patients were divided into three groups based on their Gleason score. Group 1, group 2

and group 3 consist of patients with Gleason score equal to 6, 3+4 and ≥ 4+3, respectively. A Random Forest classifier strategy was used to predict the Gleason score groups and the 5-fold cross validation strategy were implemented to evaluate its performance. More precisely, the classifier achieved an average Area Under the Curve of Receiver Operating Characteristic (AUCROC) of 83.40, 72.71 and 77.35% in predicting the group 1, group 2 and group 3, respectively. The most important radiomic features for predicting group1 were zone size percentage, large zone size emphasis and zone size non-uniformity. All these features belong to the Gray Level Size Zone matrix and also showed significant correlation with the Gleason score group after performing Kruskal-Wallis and Spearman's rank correlation tests with Hol-Bonferroni procedure for multiple corrections. Furthermore, the Entropy and the Sum Entropy features, which belong to the Gray Level Co-occurrence matrix, were the most important features for predicting group 2 and group 3, respectively.

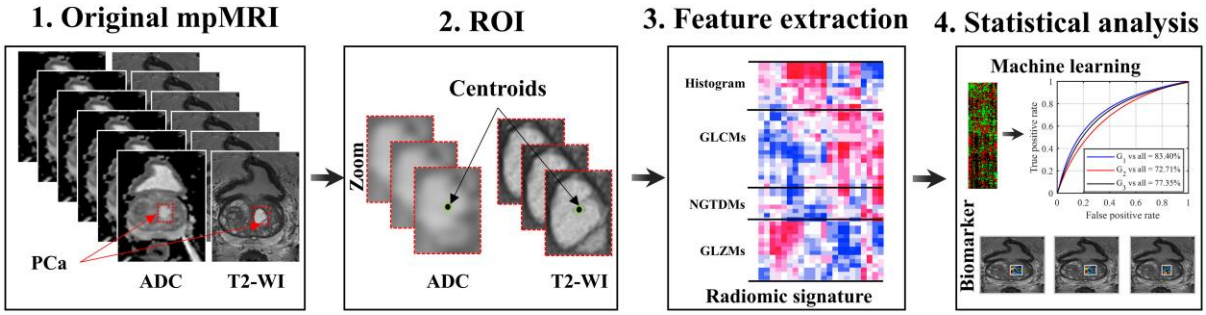

Figure 1. Flowchart for the radiomic analysis conducted by Chaddad et al. (2018) [39].

Furthermore, Li et al. [40] investigated the potential of several clinical and radiomics models for predicting clinically significant prostate cancer using bi-parametric MRI (bpMRI) (T2w and ADC). They concluded that the radiomics-based model and the combined radiomics-clinical models outperformed the clinical model, achieving AUC 98% compared to 79%, respectively. Additionally, Liu et al. [41] demonstrated that machine learning model based on radiomics extracted from Dynamic Contrast Enhanced (DCE) MRI images has also very good performance in predicting prostate cancer aggressiveness (Gleason score ≤ 7 versus Gleason score ≥ 8), achieving AUC 84% and higher.

A study by Rodrigues A. et al. [42] investigated pipelines for predicting the aggressiveness of prostate cancer using bpMRI data. Samples with Gleason score greater than 7 were considered as aggressive prostate cancer. T2w, Diffusion weighted Imaging (DWI) and ADC images from the PROSTATEx challenge were used to fit 288 different pipelines. Each pipeline was executed 50 times to account for distribution comparison, while keeping 25% of the data as a holdout dataset. The authors used segmentations from both the lesions and the whole gland for model performance comparison. In addition, they constructed four datasets using: a) the radiomics from the lesion; b) the radiomics from the whole gland; c) the lesion radiomic features and features that describe the anatomical location of the lesion; and d) whole gland radiomics from images with single lesions. In the pipelines, they included several sampling strategies, feature selection methods and machine learning algorithms (Figure 2). The results

suggested that features extracted from the whole gland were more stable than features extracted from the lesion Volume of Interest (VOI). Furthermore, features extracted from the whole gland seem to provide helpful insights for predicting the prostate cancer aggressiveness.

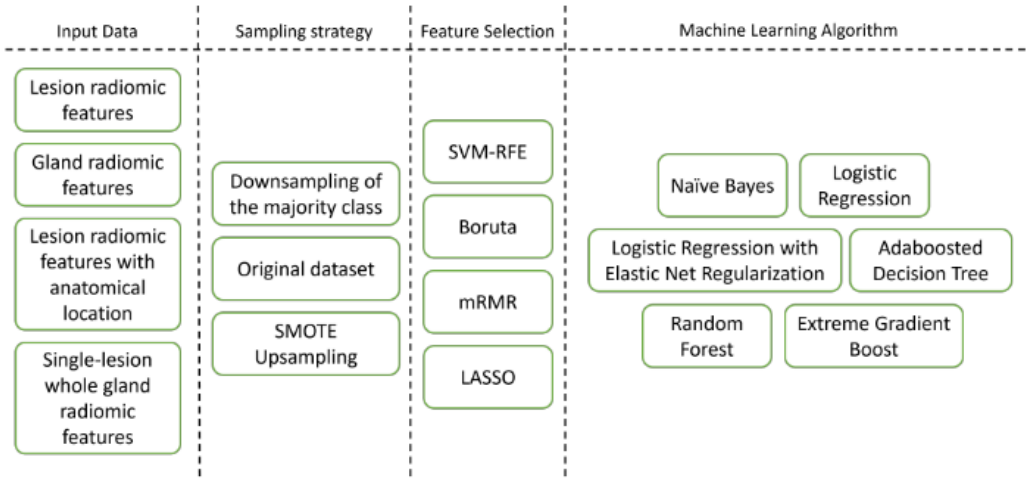

Figure 2. Pipeline configurations investigated by Rodrigues A. et al. (2021) [42].

Similar to our study, Sun P. et al. [43] investigated the predictive value of several machine learning pipelines on glioma grading. The authors fitted the 240 combinations of 16 feature selection methods and 15 classifiers with radiomic features extracted from 210 patients with glioblastoma and 75 with low-grade glioma. For each patient, images from four modalities were used, including T1 gadolinium (T1-Gd), T1, T2 and FLAIR. Cross validation and random train-test split strategies were used for evaluating the predictive performance of all the combinations. Their results were promising for glioblastoma classification, since the optimal combination achieved balanced accuracy 94.40% and AUC 98.6% using 10-fold cross validation. The balanced accuracy increased to 95.3% and an AUC of 98.1% was achieved when using the random train-test splitting strategy. Finally, the results suggested that the predictive performance of the models were affected by both the feature selection methods and the classifiers used for grading glioblastoma.

Furthermore, deep learning models have been developed to predict clinically significant prostate lesions, showing promising performance. Seetharaman et al. [44] implemented a convolutional neural network using as input T2W and ADC maps, achieving AUROC equal to 75%. Bhattacharya et al. [45] leveraged MR images and histopathology images to combine the information stemming from these two modalities. They identified correlated deep features between radiology and pathology images and fed them into a convolutional neural network to predict clinically significant lesions. The model's performance was equal to 82 and 86% in radical prostatectomy and biopsy cohort patients, respectively.

Bertelli et al. [46] conducted a monocentric study and investigated the ability of machine learning and deep learning models to predict the prostate cancer aggressiveness using T2W,

ADC and combined T2W and ADC images. The prediction of the tumor's aggressiveness was based on the ISUP score. More specifically, the patients were divided into two groups, i.e., patients with low grade (ISUP ≤ 2) and patients with high grade (ISUP ≥ 3). They utilized 2 cohorts, consisting of 85 (PI-RADS 2.0) and 27 (PI-RADS 2.1) patients, respectively. They extracted 95 radiomic features for each slice from T2W images and ADC maps and applied data augmentation techniques, such as Adaptive Synthetic (ADASYN), Synthetic Minority Oversampling Technique (SMOTE) and its variants. Ensemble classifiers were used to combine the advantages and the predictions of single classifiers to boost the final performance. Furthermore, a deep learning analysis was applied implementing Convolutional Neural Networks (CNN) on 2D data. The results showed that both the machine learning and deep learning models had better performance when trained on T2w images. More specifically, the machine learning and the deep learning model achieved an AUROC of 75% and 87.5%, respectively, when tested on the hold-out PI-RADS 2.0 test set with T2w images.

Another study conducted by Castillo et al. [47] compared the performance of deep learning and radiomics models on classifying the clinically significant prostate cancer using mpMRI (T2w, DWI and ADC). They tested their models in 3 external multicentric cohorts, consisting of 374 patients in total. The patients with ISUP grade equal or larger than 2 were classified as significant prostate cancer. The results showed that the radiomic model outperformed the deep learning model in the three independent testing sets, achieving AUCs of 88, 91 and 65% compared to 70, 73 and 44%, respectively. Thus, this study concluded that radiomic model is more generalizable and accurate model for predicting clinically significant prostate cancer than deep learning model.

Chapter 3: Research methodology

This study identifies an optimal pipeline for detecting prostate cancer aggressiveness. Deploying machine learning models for such purpose, while achieving high performance, is a challenging task. In this section, the methodology used for developing an optimal pipeline is thoroughly explained. Machine learning workflow, including the preprocessing steps, the feature extraction, standardization and feature selection methods used, is presented in the following sections. The schematic representation of the radiomics workflow used in the current study is depicted in Figure 3. The analysis was implemented using Python (version 3.9) programming language on a computer with Ubuntu 22-LTS, 16-core CPU and 64GB RAM.

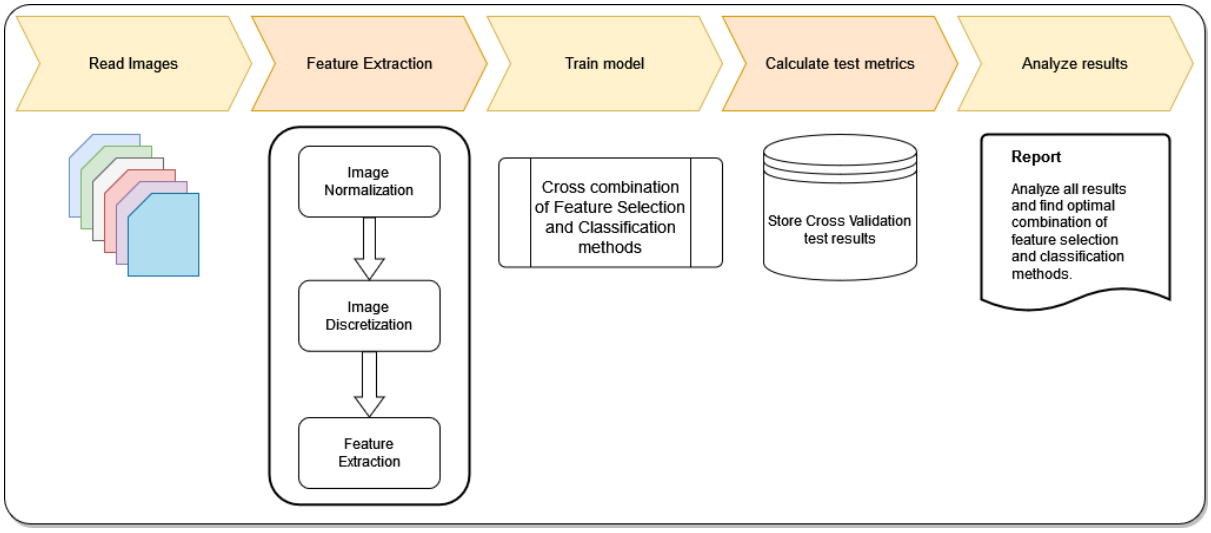

Figure 3. Radiomics workflow used in the current study

3.1 Dataset description

One of the most important components of a machine learning workflow is the dataset used in the analysis. The quantity and quality of the dataset directly affect the performance of the models. In the current study, the publicly available dataset from the "PI-CAI" (Prostate Imaging: Cancer AI) challenge [48] (accessed on February 5, 2023) is used for the development of the radiomic pipeline. It is a multi-center and multi-vendor dataset consisting of 1500 bpMRI prostate exams from 1476 patients, including annotations, as well as clinical and acquisition metadata. There are two modalities included in the dataset, i.e. T2w and DWI. The annotations are derived from DWI, but they were resampled for the T2w images and are used in the current study. Data was acquired retrospectively and provided from three Dutch centers, i.e. Radboud University Medical Center (RUMC), Ziekenhuis Groep Twente (ZGT), University Medical Center Groningen (UMCG), and one Norwegian center the Norwegian

University of Science and Technology (NTNU). Clinical data include patient and study identifiers, PSA and PSA density values, prostate volume, MRI exam date, patient age, histopathological type, the ISUP score and Gleason score.

Gleason score is a grading system for prostate cancer, developed in 1960s by Gleason D. et al. [49] and updated in 2014 [25]. This system examines the patterns of cancer cells in the prostate tissue, how they behave and look compared to normal cells. Tumor cells fall into 5 distinct patterns and are graded on a scale of 1 to 5 according to the observed behaviour and appearance. Grade 1 cells resemble to normal prostate tissue, while grade 5 cells have been mutated resulting in significantly different tissue appearance than normal prostatic tissue. These patterns are combined and form the Gleason Score. For instance, tumor cells can have two patterns indicated with a combination of two numbers, from which the first one corresponds to the most prevalent pattern. A tumor with patterns 3 and 4 with prevalence of pattern 3 corresponds to a Gleason Score of 3 + 4 = 7. Thus, patterns 3 + 4 and 4 + 3 are not the same, since in the second case the cancer is more aggressive, as pattern 4 is more prevalent among the cells. Finally, in 2014 an updated prostate cancer grading system was proposed by the International Society of Urological Pathology (ISUP), called the Grade Group (GG) or ISUP score [25]. ISUP score is strongly correlated with Gleason Score (GS) and their correspondence is shown in Table 1.

Table 1. Grade Group (ISUP score) and Gleason score correspondence

Gleason score	ISUP score	Description
6 (3 + 3)	1	Cancerous cells tend to grow slowly
7 (3 + 4)	2	Most cancer cells tend to grow slowly, while the rest grow moderately
7 (4 + 3)	3	Most cancer cells tend to grow moderately, while the rest grow slowly
8 (4 + 4)	4	All cancer cells tend to grow moderately
9 (4 + 5, 5 + 4), 10	5	Cancer cells are likely to grow moderately to quickly

From 1500 MRI scans, 220 are accompanied by manually extracted delineations, while 5 of the 220 exams have multiple lesions. Thus, there are 225 samples of csPCa, which are identified by the ISUP score evaluated by clinicians. In Figure 4, a representative sample of the dataset is depicted along with the annotated lesion. Samples are split in two cohorts based on the aggressiveness of the tumor. The first cohort contains the cases (*n=135*) where the ISUP score is equal to 2 (low aggressiveness), and the rest of the samples (*n=90*) lie in the second cohort where ISUP is greater than 2 (high aggressiveness). This splitting is based on the proliferation rate of the tumor cells, as tumors with ISUP greater than 2 tend to be more aggressive due to the rapid growth of cancerous cells [50].

Figure 4. Annotated sample of the dataset with single Region Of Interest (ROI). The ROI is depicted with green color.

The PI-CAI dataset contains images acquired using MRI scanners from Philips and Siemens, including multiple models of the two manufacturers and different magnetic field strengths. Table 2 shows the detailed distribution of the 220 images used in the current study in terms of vendor and model used for data acquisition.

Table 2. Distribution of samples per vendor

Vendor	Model	Magnetic Field Strength	# of samples per vendor	# of samples per model
Philips Medical Systems	Achieva	1.5T	83	19
	Ingenia	3T		64
Siemens	Aera	1.5T	137	6
	Prisma	3T		12
	Skyra	3T		97
	TrioTim	3T		22

Table 3 and 4 summarize the number of samples for each class per vendor and per magnetic field strength, respectively. The distribution of samples for each class per vendor is suitable for stratifying

the data, since it is more balanced than the distribution based on the field strength and it efficiently captures the variability between the vendors.

Table 3. Distribution of images per vendor

	Philips Medical Systems	Siemens
Low aggressiveness csPCa	56	79
High aggressiveness csPCa	27	63

Table 4. Distribution of images per magnetic field strength

	1.5T	3T
Low aggressiveness csPCa	18	117
High aggressiveness csPCa	7	83

3.2 Image preprocessing

Several preprocessing steps are used in radiomic studies to improve the quality of the data and decrease the variability in the image intensities. Especially in multicentric studies, the differences in the vendors and acquisition protocols may result in significant variations in the image intensities. In order to reduce the inconsistencies and the variability, image normalization techniques and a bias field correction method were used in the current study to assess their impact on the model's performance for the PCa aggressiveness.

3.2.1 Bias field correction

MR images suffer from a low-frequency variation in their acquired signal, resulting in intensity inhomogeneities. This non-uniformity is called bias field and is generated due to poor radiofrequency coils, gradient eddy currents, variations in flip angle and subject-scanner interactions. Bias field correction methods are categorized into prospective and retrospective methods [51]. The former calibrate and improve the acquisition process in order to remove the bias field. The latter aim to reduce the bias field generated by the properties of the scanned object and are more frequently used. The retrospective methods are divided into the following four categories: a) filtering methods; b) surface fitting-based methods; c) intensity-based methods; and d) histogram-based methods.

The N4ITK bias field correction method [8] is a retrospective histogram-based technique and has been very widely used as a preprocessing step in radiomic studies [52], [53]. The N4ITK method is the state-of-the-art method for bias field correction and is an improvement of the N3 filter [54]. The N4ITK filter is available in python by the open-source toolkit SimpleITK. The improvements of the N4ITK are the multi-resolution B-spline fitting routine and the optimized

iterative process. The algorithm is an iterative process of deconvolving the histogram by a Gaussian, estimating the corrected intensities and smoothing the bias field using the B-spline model. The N4ITK has several parameters that their values should be defined. These parameters are the convergence threshold, the shrink factor, the fitting level, the number of iterations and the use of a mask. The convergence threshold is the stopping criterion of the iterative process and the shrink factor defines how much the original sample will be downsampled before estimating the bias field. The fitting level defines the number of levels that will be used to determine the resolution of the B-spline grid and the number of iterations refer to the number of the maximum iterations at each level. In this study, we used parameter values that have been identified as optimal for prostate images scanned using a surface coil by Dovrou et al. [33]. More specifically, they identified that the optimal configuration of the N4ITK filter for these images of PI-CAI dataset is: convergence threshold 0.001, shrink factor 2, fitting level 5, number of iterations 25 and the use of default mask (use of non-zero values of the image). Thus, all the MR prostate images were bias field corrected using this optimal configuration of the N4ITK filter in order to reduce the intensity inhomogeneities.

3.2.2 Normalization methods

Furthermore, intensity normalization techniques were applied to the images in order to reduce the variability in the intensity values of the images. More precisely, three normalization methods were applied. The state-of-the-art Z-score normalization method was used, which rescales and shifts the standardized intensities by the mean value of the signal intensities of the image. The pixel values are normalized according to the following equation:

$$Inew(x) = \frac{I(x) - \mu}{\sigma}$$

Where $I_{new}(x)$ is the normalized value of pixel x, $I(x)$ is the original value of pixel x, μ is the mean value of the signal intensities of the image and σ is the standard deviation of the signal intensities of the image.

Moreover, two biologically-motivated normalization techniques were used in order to homogenize the signal intensity space. These methods are based on the concept of White Stripe normalization [55] method, which was developed for normalizing brain images. To this end, two pelvis specific methods were applied to the MR prostate images, called fat-based normalization and muscle-based normalization technique. In order to apply these methods, an approximation of the fat and the muscle tissue should be identified. To this end, a segmentation method was used to automatically segment the fat and muscle tissue in MR pelvic images. Firstly, the N4 bias field correction method was applied to the images in order to produce images free from bias field artifacts. Each image was subsequently cropped by removing the 20% of the columns in the middle of the image to automatically remove the heterogeneous prostate gland and simultaneously maintain the largest area of the fat and the muscle tissue. The fat signal intensity is expected to be the highest among the other abundant

tissues of the pelvic region in T2W imaging of the prostate. As opposed to the fat tissue, muscle tissue is expected to occupy the low signal distribution of the spectrum in the whole histogram. The K-means algorithm was applied to the cropped image, setting K equal to 2, in order to identify the 2 clusters of the low intensity values (i.e. muscle tissue approximation) and the high intensity values (i.e. fat tissue approximation). Especially for the segmentation of the muscle tissue, the 12th percentile of the distribution is calculated in order to remove the 12% of the lower values that correspond to background pixels representing air and the vessels. The effect of this percentile was assessed by an experienced radiophysicist evaluating the results obtained by different percentiles. An example of an N4 filtered image after removing the 20% in the middle and the corresponding fat and muscle segmentations (with red color) are presented in Figure 5.

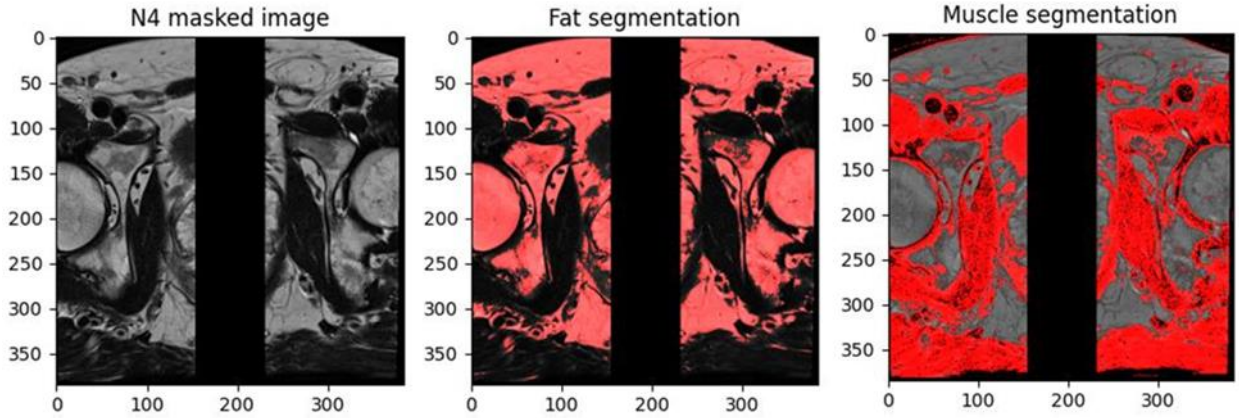

Figure 5. From left to right, the N4 filtered image after cropping the 20% of the image in the middle, the fat segmentation (with red color) and the muscle segmentation (with red color) are presented.

In the fat-based normalization method, the fat tissue, which was automatically segmented, was used as reference tissue in order to normalize the whole image according to statistics derived from the fat's distribution. More precisely, in the fat-based normalization method, the image intensity values are transformed according to the following equation:

$$I_{fat-normalized}(x) = \frac{I(x) - \mu_{fat}}{\sigma_{fat}}$$

where μ_{fat} is the mean intensity value of the voxels that correspond to the fat tissue and σ_{fat} is the standard deviation of the voxels that correspond to the fat tissue.

Accordingly, in the muscle-based normalization technique, the image intensity values are transformed according to the following equation:

$$I_{muscle-normalized}(x) = \frac{I(x) - \mu_{muscle}}{\sigma_{muscle}}$$

where μ_{muscle} is the mean intensity value of the voxels that correspond to the muscle tissue and σ_{muscle} is the standard deviation of the voxels that correspond to the muscle tissue.

Hence, the aforementioned bias field correction and normalization methods were applied independently and in combination to the MR prostate images in order to assess the effect of each preprocessing pipeline to the model's performance. More precisely, six different datasets were derived and used for subsequent analysis, which are: a) original dataset; b) Z-score normalized dataset; c) N4 bias field corrected dataset; d) N4 bias field corrected and Z-score normalized dataset; e) fat-based normalized dataset; and f) muscle-based normalized dataset. The 6 datasets used in the analysis and their brief description are presented in Table 5.

Table 5 Descriptions of the datasets used for the analysis

Dataset	Description
Original	The PI-CAI dataset described in section 3.1
Original Normalized	Z-score normalized dataset
N4	N4 filtered dataset
N4 Normalized	N4 filtered dataset with Z-score normalization
Fat	Fat-based normalized dataset
Muscle	Muscle-based normalized dataset

3.3 Radiomics extraction

Radiomic feature extraction requires several parameters, which configure the extraction process, the number and type of features. An important parameter is the width of the bins that will be used to discretize images before extraction. Bin width was calculated using a fixed bin size and the mean range of intensities per image, using the following equation:

$$BinWidth = \frac{MeanRange}{BinCount}$$

For the purposes of this study, $BinCount$ is set to 32 and ranges are calculated by loading every image, subtracting the minimum value from the maximum value and then calculating the average range for a specific dataset. Using the above equation, the bin width is calculated for the original and all preprocessed datasets for extracting radiomic features. Finally, all images were resampled to isotropic voxel size of 1mm using the B-Spline interpolator.

In this study, features were extracted with the Python library *pyradiomics* [56], which is commonly used in the literature for radiomic feature extraction. A total of 1132 features were automatically extracted from the segmented region of each image using: i) the original image without any filtering, ii) wavelet filtered images and iii) Laplacian of Gaussian filtered images with sigma values 2, 3, 4, and 5. The extracted features are related to the distribution of intensity levels and they were calculated using the histogram (first order statistics), shape and texture, i.e. Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM) and Gray Level Dependence Matrix (GLDM). Extracted features for all patients are saved in comma separated values (csv) files, which additionally include the patients' ids and the assigned ISUP score per lesion. All the features were standardized as an initial preparatory step for subsequent analysis.

3.4 Feature Selection Methods

Data with high dimensionality are difficult to be handled, as models have the tendency to overfit on large feature spaces, a phenomenon known as the "Curse of Dimensionality" (CoD). Hence, CoD causes lagging performance on unseen data, and more sophisticated models are required for achieving high accuracy [9]. Thus, after the feature extraction process, it is necessary to select a subset of those features to reduce the computational costs and improve models' performance. The selection process is not trivial and should be meticulously conducted in order to keep only the most informative and not redundant features. There are several feature selection techniques used in the literature, a subset of which is used in the current analysis. The 13 feature selection methods used in this study are: Pearson correlation, Spearman correlation, Minimum Redundancy Maximum Relevance (mRMR), Mutual Information Feature selection (MIFS), Conditional Mutual Information Maximization (CMIM), Joint Mutual Information (JMI), Boruta, Least Absolute Shrinkage and Selection Operator

(LASSO) and Relief family of algorithms. In order to use these feature selection methods, the *ITMO_FS* (univariate and multivariate filtering methods), *sklearn* (LASSO), *Boruta_Py* (Boruta) and *skrebate* (RBAs) libraries were used. The LASSO feature selection method and methods from *ITMO_FS* were modified to be compatible with sklearn library, following the library's guidelines[1]. Thus, these methods can subsequently be used in the pipeline. The implementation of each method is briefly described.

3.4.1 Pearson and Spearman correlation

Both Pearson and Spearman feature selection are univariate filter methods, which select features based on the correlation between two variables. Thus, the correlation between each feature and the target variable or between two features can be calculated. These techniques are quick, but they do not take into consideration the interactions between more than two variables. Thus, it is not guaranteed that only non-redundant features are selected and may hamper the prediction accuracy. In the current study, Pearson and Spearman correlations are calculated between all possible pairs of features. For every pair that exhibits a correlation higher than a predefined threshold, the feature with the greatest average correlation among all features is eliminated. Thus, an initial filtering of the highly correlated features is performed in order to face the multi-collinearity.

Pearson correlation coefficient measures the linear association between two variables, using the following formula:

$$corr_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where σ_{XY} is the covariance between X and Y, σ_X the standard deviation of X and σ_Y the standard deviation of Y. The values of Pearson coefficient are from -1 (fully negative correlation) to 1 (fully positive correlation), while $corr_{X,Y} = 0$ indicates no correlation between the two variables. In addition, a positive correlation means that when the value of one variable increases, then the value of the other variable also increases. In contrast, a negative correlation shows that when the value of the one variable increases, then the other decreases. Spearman correlation coefficient measures the non-linear associations between the features. It works in the same way as the Pearson correlation, but it uses their ranks instead of calculating the covariance and standard deviation directly on the features. Both methods are widely used for feature selection, since they greatly reduce the initial feature space.

[1] For the technical guidelines used see: https://scikit-learn.org/stable/developers/develop.html

3.4.2 minimum Redundancy Maximum Relevance (mRMR)

Minimum Redundancy Maximum Relevance (mRMR) is one of the most commonly used feature selection methods in the literature [35]. It is a multivariate filter method which keeps features that are informative for the prediction task and avoids redundancy. The selection process begins with an empty feature set and the feature that is highly relevant with the target variable is selected. For determining the most informative feature, the f-statistic is calculated. The next feature selected is the feature that is highly relevant with the target variable and has simultaneously minimum redundancy with respect to the selected feature. Redundancy is calculated as the average Pearson coefficient for all selected features. This process iteratively selects the feature with the maximum feature score, until the feature set size reaches the predefined limit.

In each iteration, the i-th feature score is calculated as:

$$FeatureScore_i = \frac{relevance(f_i, target)}{redundancy(f_i, SelectedFeaturesSet)}$$

where $SelectedFeaturesSet$ is the set of selected features, $relevance$ is the f-statistic between the feature and the target label and redundancy is calculated as:

$$redundancy(f_i, SelectedFeaturesSet) = \frac{1}{M}\sum_{s \in SelectedFeaturesSet} |PearsonCorr(s, f_i)|$$

where M equal to the number of already selected features and $PearsonCorr$ is the Pearson correlation coefficient between the feature f_i and the selected feature s. Note that the absolute value of Pearson correlation coefficient is used, as the magnitude of the correlation is important rather than the sign (positive or negative).

3.4.3 MIFS, CMIM, JMI methods

Mutual Information Feature selection (MIFS), Conditional Mutual Information Maximization (CMIM) and Joint Mutual Information (JMI) methods are also multivariate filters that not only select the most informative features for the target variable, but they also keep features that are not highly correlated with each other. These methods work iteratively like the mRMR method but they use different measures for selecting features. They are all based on the entropy decrease that occurs if a feature is selected. Entropy is the randomness that exists in a system and thus the features should be selected in a way that the total entropy of the system is reduced.

MIFS method uses mutual information (I), which is the randomness that is removed from the system when a feature is selected, providing more information for the target variable. Moreover, this method introduces a penalty for reducing feature redundancy. The feature score is calculated by the following equation:

$$FeatureScore_{MIFS}(f_i) = I(f_i; target) - \beta \sum_{s \in SelectedFeaturesSet} I(f_i; s)$$

The penalty term is multiplied by a coefficient β which was found through experiments that its optimal value is 1, without strong proof [57].

JMI method uses the joint mutual information criterion for selecting features. The basic idea is that every feature that is 'complementary' with the already selected features should be included. The feature score is calculated as:

$$FeatureScore_{JMI}(f_i) = \sum_{s \in SelectedFeaturesSet} I(f_i\, s; target)$$

Finally, CMIM method maximizes the conditional mutual information criterion, which includes features that are informative with respect to target, while simultaneously considers the conditional relations between features. This method initially selects the feature with the highest mutual information with the target variable. Then, it iteratively selects features based on the conditional mutual information with the target and the relevant information gained with the already selected features. The feature score can be evaluated using the following equation:

$$FeatureScore_{CMIM}(f_i) = I(f_i;\ target) - \max_{s \in SelectedFeaturesSet} [I(f_i;\ s) - I(f_i; s|target)]$$

For more information about these methods see [58]. Furthermore, these methods provide feature sets that are more likely to improve the accuracy of the models, but the calculations are slower than the univariate filtering methods.

3.4.4 Boruta

Boruta is a wrapper feature selection method that leverages the Random Forest (RF) classifier for generating feature sets, containing all the significant features [37]. It generates randomized permutations of the features (shadow features) and selects the features performing better than the best shadow feature. The performance measure is the feature importance that is inherently generated by the RF algorithm. After the first iteration Boruta finds all features that had greater importance than a specific threshold, called hits. This threshold is equal to the maximum importance of all the shadow features. A statistical two-sided equality test for all features is used to determine whether the feature is accepted or rejected. This iteratively process of selecting features continues for a predefined number of iterations or until all features are accepted or rejected. Moreover, Boruta algorithm may terminate with some features that are not accepted, neither rejected; thus, the algorithm is indecisive about those features (weak features) and the machine learning developer should decide whether these features will be included. In the current study, the weak features are excluded from the subsequent analysis. To conclude, Boruta is an efficient feature selection

technique based on a simple statistical test that additionally provides an overall ranking of the features.

3.4.5 Least Absolute Shrinkage and Selection Operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO) feature selection is another commonly used algorithm [36]. It is an embedded method based on Lasso regression, which is a linear model that adds a scalable penalty term to the least squares cost function. The complete cost function with the l_1 penalty is:

$$\frac{1}{2N}\sum_{i=1}^{N}\left(y_{real}^{(i)} - y_{pred}^{(i)}\right) + \alpha\sum_{j=1}^{n}|a_j|$$

where N is the number of training samples, a_j the coefficient of j-th feature, n the number of features and the hyperparameter α scales the penalty term. If the lasso regressor discovers two features which are linearly correlated, it will attempt to shrink the coefficient of the less important feature to 0, for optimizing the cost function. After the optimization is completed, the features with coefficients equal to zero are discarded; thus, only the important and non-redundant features are preserved. Standardization of the data before training the lasso regression model and tuning of the hyperparameter α using cross validation (CV) are two mandatory steps. Thus, for each examined dataset, the value of the hyperparameter α was investigated using 3-fold cross validation.

3.4.6 Relief family of algorithms

The algorithms of the Relief family estimate the quality of the features based on their ability to distinguish instances that are near to each other. They are able to discover any strong dependency between the features, while correctly estimating the features' quality.

The basic Relief algorithm [59], [60] is used when feature selection is applied on a two-class classification task. Initially all features' weights are equal to zero. In every iteration, the algorithm selects a random instance (I), finds the nearest samples of the same class (hit) and the nearest sample of the opposite class (miss) and it updates the weights according to the following equation:

$$W[f_i] = W[f_i] - \frac{diff(f_i, I, H)}{m} + \frac{diff(f_i, I, M)}{m}$$

where H is the nearest hit, M the nearest miss, m the number of total iterations and the $diff$ function is defined based on the features category. For categorical features, $diff$ function is defined as:

$$diff(f_i, I_1, I_2) = \begin{cases} 0; value\ of\ feature\ f_i\ is\ equal\ in\ both\ instances\ I_1, I_2 \\ 1; otherwise \end{cases}$$

while for numerical features is defined as:

$$diff(f_i, I_1, I_2) = \frac{|value\ of\ f_i\ in\ I_1 - \ value\ of\ f_i\ in\ I_2|}{\max(f_i) - \min(f_i)}$$

If the distance between the randomly selected instance of a positive class and the hit (which is also in the positive class) is large, or the distance of the selected instance and the nearest miss (which is in the negative class) is small, then the feature separates two instances of the same class (positive class) and it does not separate samples of different class. This behavior is not desirable and thus the algorithm reduces the weight of this feature. Thus, the algorithm tries to find high quality features that best discriminate samples of different classes.

In this study, several Relief-based algorithms (RBAs) were used as feature selection methods. An extended version of the Relief algorithm, called ReliefF, is used, which handles multiclass classification tasks and missing data. In this variant, the algorithm finds the k-nearest hits and misses, instead of only one. After randomly selecting the instance I, it finds the k-nearest hits and for every class different than the class of the selected instance, it finds the k-nearest misses. The weights are updated as follows:

$$W[f_i] = W[f_i] - \sum_{j=1}^{k} \frac{diff(f_i, I, H_j)}{m * k} + \sum_{C \neq class\ of\ I} \left[\frac{P(C)}{1 - P(class\ of\ I)} \sum_{j=1}^{k} \frac{diff\left(f_i, I, M_j(C)\right)}{m * k} \right]$$

where k is the number of nearest neighbors, $P(C)$is the probability of the class C, and $M_j(C)$ is the j-th miss that belongs to class C. As the parameter m approaches the number of instances n, the weight estimations are getting more reliable [61]. The $diff(f_i, I_1, I_2)$ function is also updated for probabilistically managing missing data. If one of the instances (e.g. I_1) has missing data, the $diff$ function become:

$$diff(f_i, I_1, I_2) = 1 - P(value\ of\ f_i\ in\ the\ instance\ I_2 | class\ of\ I_1)$$

If both instances have missing data, then the $diff$ function become:

$$diff(f_i, I_1, I_2) = 1 - \sum_{V}^{\#\ of\ values\ of\ f_i} (P(V\ | class\ of\ I_1) \times P(V\ | class\ of\ I_2))$$

For more information about the implementation of Relief and ReliefF algorithms see [62].

In addition to ReliefF, other variants of Relief algorithm, i.e. Surf, Surf Star, Multi Surf and Multi Surf Star, which are implemented in the *skrebate* [63] Python library, are also used in this study. All these variants have the same core idea with the ReliefF, but differ in terms of neighbor selection and weights updating.

Surf [64] algorithm uses the notion of threshold-based neighbors, where all instances that are in a distance less than a threshold T are considered neighbors and are weighted equally. The value of T is set equal to the average pairwise distance between all instances. The extended version Surf* (Surf star) [65] uses the same threshold T, introducing the "far" scoring. All instances that are within the T are considered hits and those outside T or "far" from the current instance are considered misses. Also, the algorithm weights differently each neighboring instance. The hits are weighted with $n + 1$, where n is the number of features, decreasing the feature score and misses are weighted with $n - 1$, which yields an increase in the feature score, respectively.

Moreover, the Multisurf* (Multisurf star) [66] algorithm introduced a dead-band-zone, where all instances in this zone have zero weights. The limits for this zone are $T_{near} = T + \sigma$ and $T_{far} = T - \sigma$, where T is a decision threshold for finding neighbors that is equal to the pairwise mean distance between the selected instance and all the others, instead of the pairwise mean distance between all instances used in Surf. The parameter σ is equal to the standard deviation of the pairwise distance between the selected instance and all the others. Hits yield an increase in feature score and misses a decrease, as opposed to the Surf* algorithm. This happens since in "far" (distance > T_{far}) instances it is more frequent to find different values.

The last algorithm used for feature selection is Multisurf [63], which applies the same logic as Multisurf*, without utilizing the notion of "far" scoring. Despite the improved ability to determine 2-way interactions, "far" scoring may fail to identify main effect interactions [38]. Multisurf is a feature selection method that allows the detection of these main effect interactions, while it is more computationally efficient than the others and can be applied to a variety of data types.

In Figure 6, illustrations depicting the selection process and weighting of neighbors in the various Relief-based algorithms are presented.

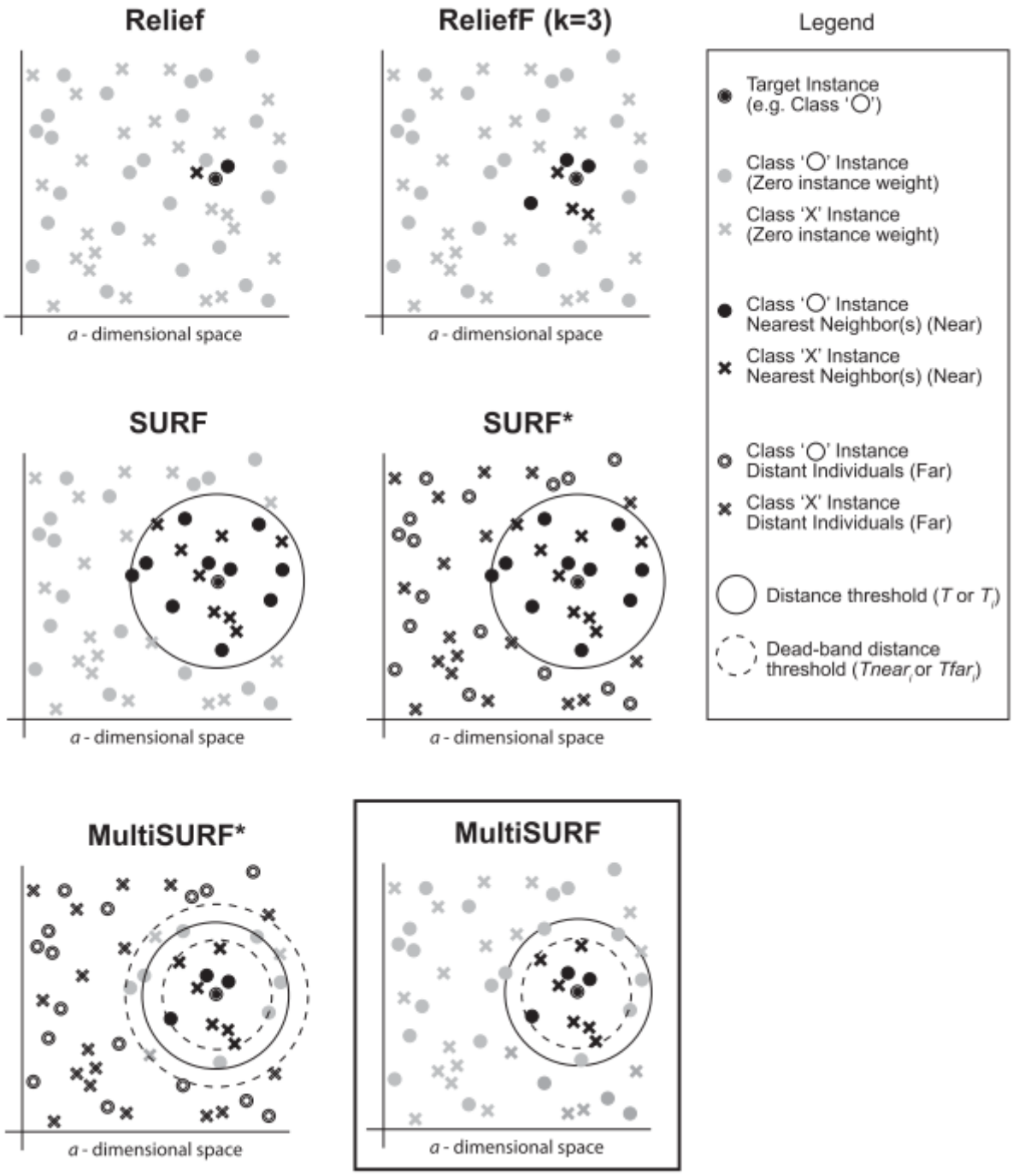

Figure 6. Illustration of how the neighbors are selected and weighted in various Relief-based algorithms adopted from [63].

3.5 Machine Learning Analysis

The next step after radiomic features extraction and selection is the classification of the samples into low aggressive ($ISUP = 2$) and high aggressive ($ISUP > 2$) csPCa. For this task, several classification algorithms are used in the literature. The most commonly used are the Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), k-Nearest Neighbour (kNN), and Extreme Gradient Boost (XGB), which are also used in this study. All algorithms are initialised with the default parameters and SVM is used with both linear and gaussian (RBF) kernel. The samples were split into a train and test set, where the test size is equal to the 40% of the total samples. A stratification strategy based on the target variable and the manufacturer of the scanner was used for splitting the data. Thus, the training set consists of 135 samples, while the test set consists of 90 samples. For every dataset and execution of the pipeline, the train and test indices are preserved for making the results comparable.

A cross-combination strategy was used in order to derive all the possible combinations of feature selection algorithms and classifiers for each dataset. Grid Search Cross-Validation (GSCV) with 4 folds was used for optimizing the feature selection method regarding its hyperparameter, while extracting the optimal feature set, on the training set. The hyperparameter for most feature selection methods is the number of selected features. More specifically, for each feature selection method that requires a specific number of features to be selected, we experimented on the number of the selected features on the cross-validation schema, using all the possible values between 3 and 100 with a step 5. For the univariate feature selection methods, the optimal threshold for excluding highly correlated features was investigated using all the possible values between 0.70 and 0.95 with a step 0.05. Furthermore, the value of the hyperparameter α of the LASSO method was investigated using the 4-fold cross validation for each dataset. The various combinations of feature selection methods and classifiers were investigated for each derived dataset.

To this end, a dynamic pipeline was developed to analyse how the various aforementioned feature selection and image preprocessing methods influence the performance of the classifiers. The analysis consists of a z-score feature normalization and dynamically injection of feature selection method and classifier, resulting in 78 possible pipelines for each dataset. The z-score normalisation was performed using the *sklearn* Python library [67].

Finally, the performance of each pipeline (i.e., specific feature selection method coupled with a specific classifier for each dataset) was assessed on the hold-out test set. More precisely, after GSCV the model is retrained on the train set using the derived optimal feature number for the examined feature selection method. The performance of the model is assessed on the test data using several metrics, which are the Area Under Curve (AUC), Accuracy, Balanced Accuracy, F1-score, Precision, Recall, and Cohen's Kappa. The balanced accuracy, the F1-score and the Cohen's Kappa are useful metrics when assessing the performance of a classifier on imbalanced dataset. The Confusion Matrix was also calculated. The model assessment methodology is shown in Figure 7.

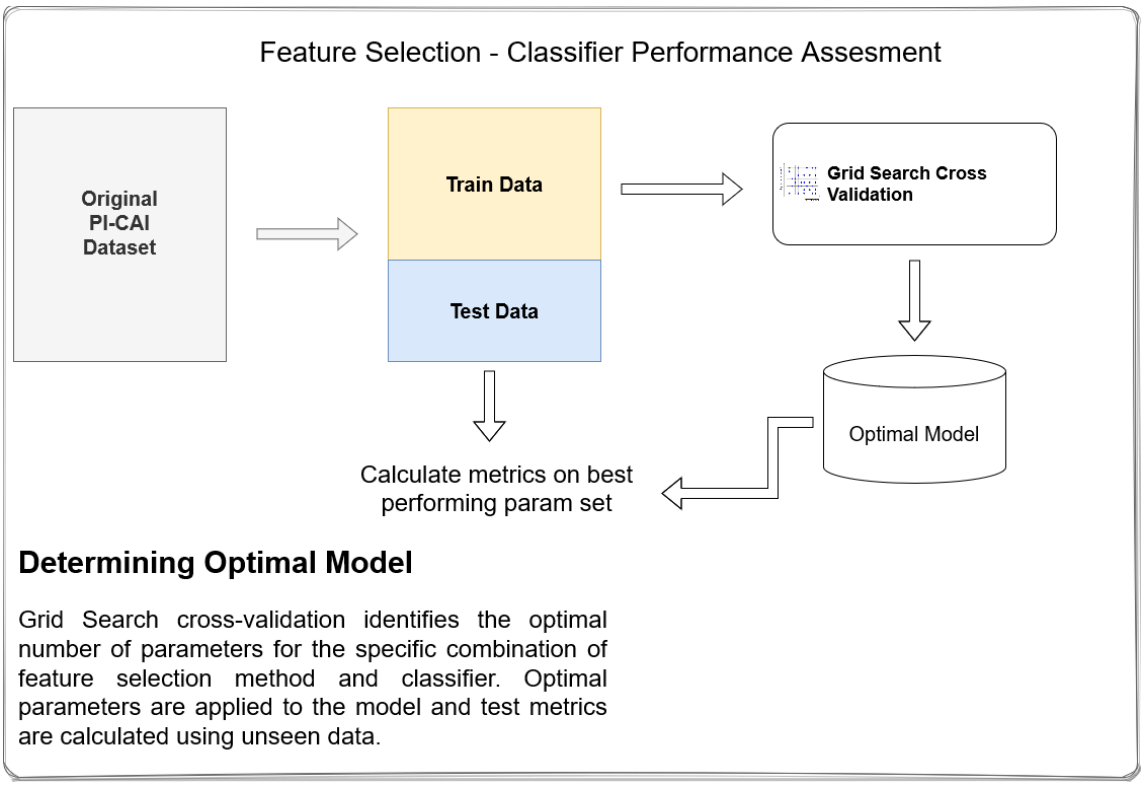

Figure 7. Main analysis workflow

3.6 Hybrid Feature Selection

After determining the optimal feature number for every combination of classifier and feature selection method, the study further investigates the impact of using a combination of feature selection methods on the performance of the models. Thus, a univariate feature selection method coupled with the optimal model extracted from the previous experiment was used. The best performing threshold in the initial analysis is used as threshold for the univariate feature selection method in the hybrid analysis. Furthermore, for the sake of completeness, GSCV is also used in the second feature selection method. The pipeline is fitted using the training data for extracting the new optimal features sets, while metrics are calculated on the unseen testing set. Figure 8 presents the pipeline.

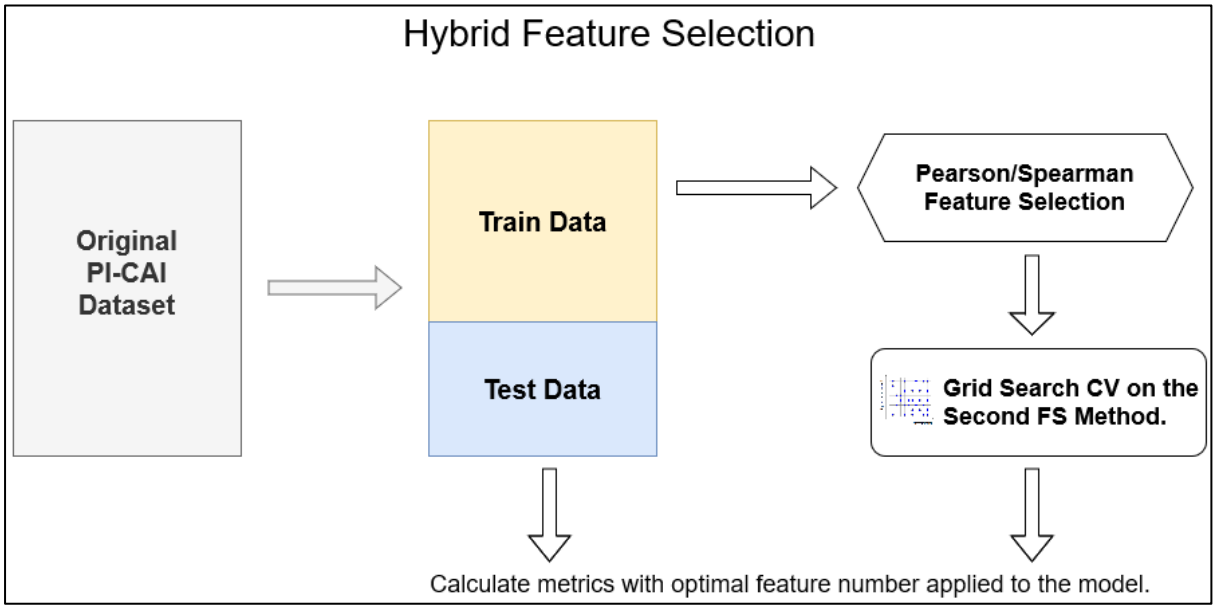

Figure 8. Analysis workflow using hybrid feature selection

Chapter 4: Research findings / results

4.1 Results of main analysis

The evaluation metrics calculated on the hold-out test set for each cross-combination pipeline (i.e. feature selection method and classifier for a specific dataset) were saved in JSON files. The results were post-processed in order to extract the optimal feature set and the model which can better discriminate the low aggressive from the high aggressive csPCa. The distributions of balanced accuracy and AUC score metrics achieved using the various pipelines per dataset are calculated and presented in the boxplots shown in Figures 9 and 10, respectively.

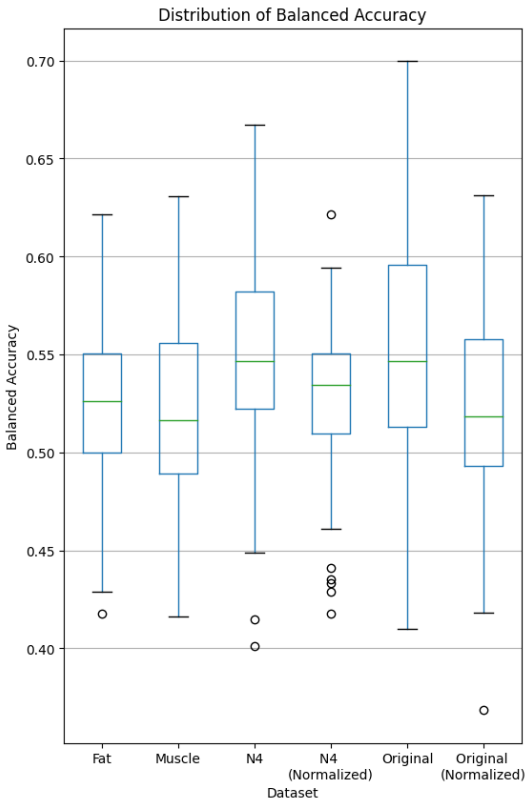

Figure 9. Distribution of balanced accuracy per dataset

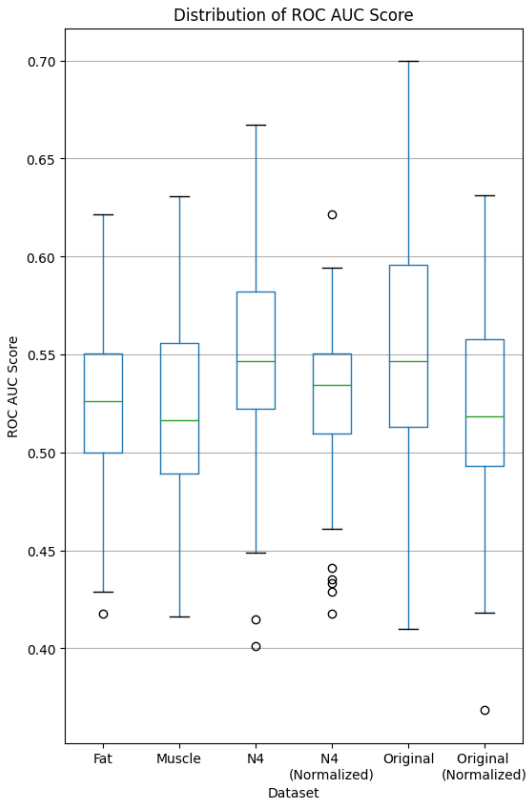

Figure 10. Distribution of AUC per dataset

Boxplots indicate that the median balanced accuracy (i.e., green line in the figure) and AUC in all datasets is in the range of 52% to 55% and of 0.52 to 0.55, respectively. The classifiers achieved higher performance using the original and N4 filtered datasets, without normalisation, than the other datasets. Moreover, the original dataset demonstrated the best performance reaching balanced accuracy equal to 70% (AUC: 0.7, precision: 65.62%, recall: 60%). The rest of the datasets demonstrated comparatively limited balanced accuracy as the normalised original, muscle-based normalised and fat-based normalised datasets achieved maximum balanced accuracy in the range of 60% to 65% (AUC: 0.62 – 0.63, precision: 48.57% - 55.55%, recall: 44.11% - 57.14%). The normalised N4 filtered dataset demonstrated poorer

performance with balanced accuracy equal to 62.16% (AUC: 0.62, precision: 48.57%, recall: 54.83%). Thus, the boxplots indicate that most combinations of feature selection methods and classifiers resulted in poor performance in each dataset as the 75^{th} percentile of the distributions of balanced accuracy and AUC is lower than 60%, reflecting the challenging nature of predicting prostate cancer aggressiveness.

The balanced accuracy is used as main metric to assess and visualize the performance of the various pipelines, as it better reflects the accuracy of the model in imbalanced datasets. In Figure 11, the balanced accuracies achieved by each classifier per feature selection method and dataset are presented. The XGB classifier performed optimally when trained and tested on the original dataset, using the Pearson or Spearman univariate feature selection method, achieving balanced accuracy of 70%. This classifier also achieved balanced accuracy higher than 60% on the N4 dataset using the JMI, LASSO, ReliefF, Surf*, mRMR and CMIM feature selection methods, with the latter achieving the highest accuracy among these methods. Gaussian Naïve Bayes classifier achieved similar maximum balanced accuracy on all datasets (less than 60%), except for the original unnormalized and normalized dataset. More precisely, using Surf* and Multisurf on the unnormalized and the normalised original dataset, respectively, achieved balanced accuracy higher than 60%.

The kNN algorithm achieved the highest performance (balanced accuracy = 66.70%) when coupled with the ReliefF feature selection method on the N4 filtered dataset. Random Forest demonstrated optimal balanced accuracy higher than 60% using mRMR in all datasets, except for the N4 filtered and the unnormalized original dataset. The Multisurf and Spearman methods achieved the optimal performance in the N4 filtered and the unnormalized dataset, respectively. However, the best performing feature selection method for the Random Forest classifier was the Spearman rank correlation coefficient in the original dataset. The SVM classifier with the linear kernel exhibits balanced accuracy near 60% using Pearson, Spearman and ReliefF methods in several datasets. However, the use of the LASSO method increases the metric to 65.25% on the original dataset. Finally, the SVM classifier using the gaussian (RBF) kernel demonstrated a balanced accuracy larger than 60% using the Multisurf and the Surf* methods. More precisely, this classifier coupled with the Multisurf method achieved balanced accuracy of 61.69% on the fat-based normalization method. Furthermore, the use of Multisurf and Surf* methods with this classifier achieved optimal performance, reaching a balanced accuracy of 63.09% and 61.39% on the muscle-based normalization dataset, respectively.

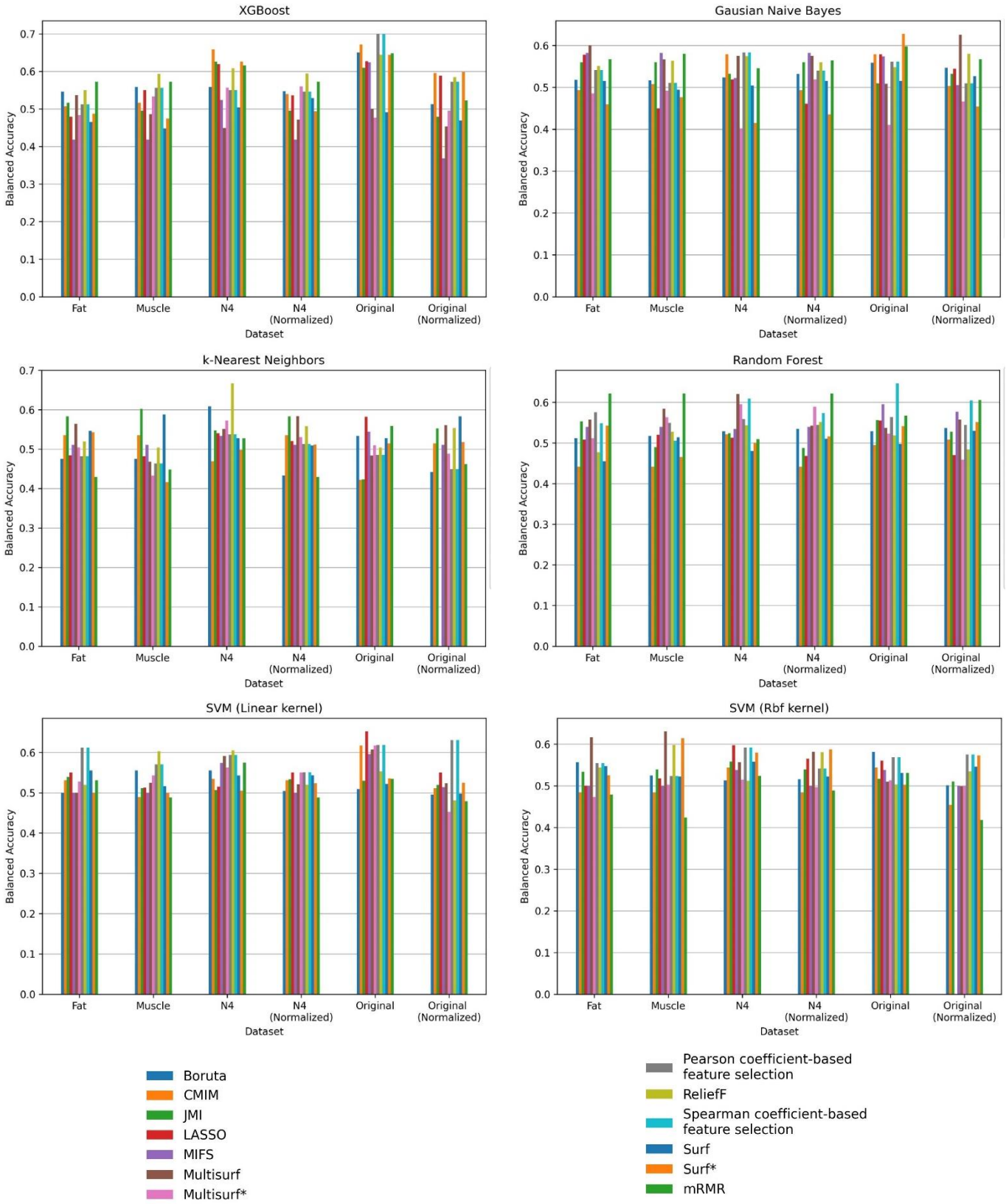

Figure 11. Balanced Accuracy for each classifier per feature selection method and dataset.

In Figures 12-14, the balanced accuracies achieved by each feature selection method per classifier and dataset are presented. Univariate feature selection methods demonstrated the best performance achieving 70% balanced accuracy on the original dataset using the XGB classifier (Figure 12). Both Pearson and Spearman correlation coefficient test had the same performance using threshold equal to 0.85. LASSO and Boruta coupled with SVM linear kernel and XGB, respectively, demonstrated a balanced accuracy of 65.25% and 64.99%, respectively, on the original dataset. All multivariate filtering methods also provided mediocre results. The mRMR method achieved a balanced accuracy close to 60% for each dataset using either Random Forest classifier or XGB (Figure 13). MIFS, JMI and CMIM have also unstable performance across all datasets with a balanced accuracy in the range of 40% to 65%. These feature selection methods achieved the best performance having balanced accuracy greater than 60% using the XGB classifier.

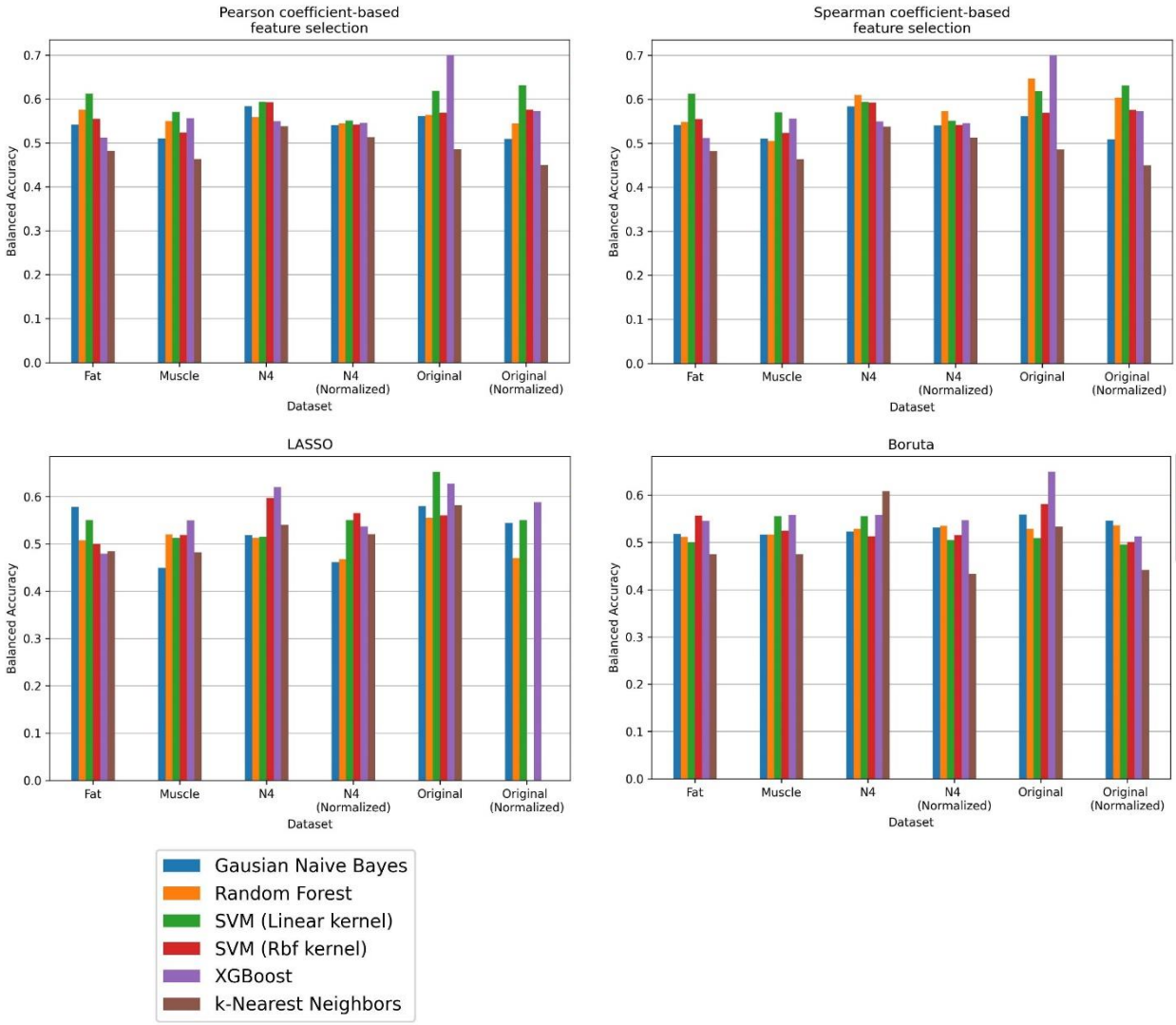

Figure 12. Balanced Accuracy for univariate, Boruta and LASSO feature selection methods per classifier and dataset.

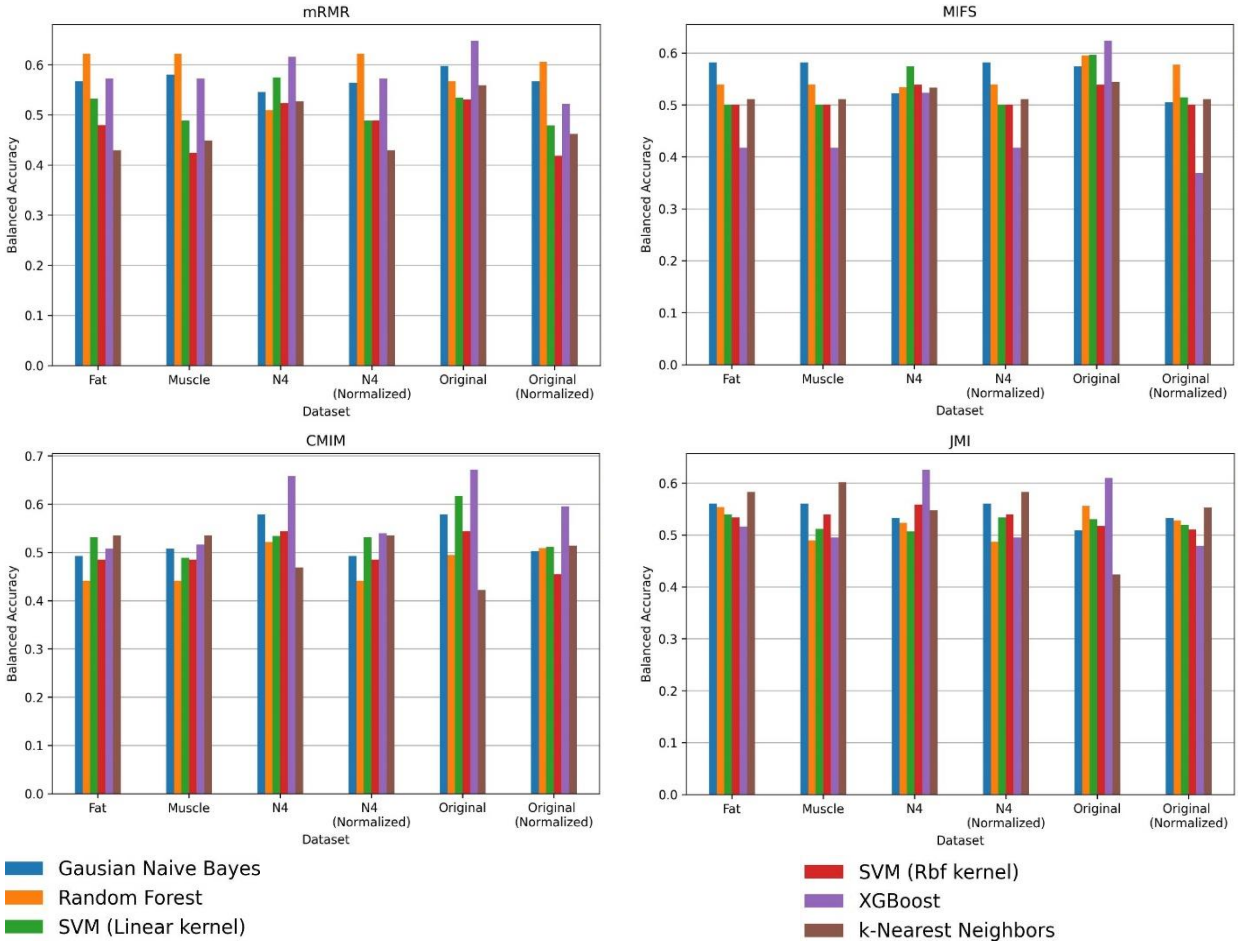

Figure 13.Balanced accuracy for multivariate feature selection methods per classifier and dataset

In addition, Relief-based algorithms demonstrated similar performance to the rest feature selection methods (Figure 14). More specifically, ReliefF achieved an optimal balanced accuracy of 66.70% when using the kNN classifier on the N4 filtered dataset. None of the classifier achieved a balanced accuracy greater than 60% when using the Surf method as feature selection technique. In contrast, the SVM with rbf (gaussian) kernel, the XGB and the Gaussian Naïve Bayes resulted in a balanced accuracy greater than 60% using the Surf* method on the muscle, N4 filtered and original dataset, respectively. Furthermore, the use of Multisurf method resulted in a balanced accuracy larger than 60% in several classifiers in all datasets, except for the N4 normalized dataset. The Multisurf* method achieved a balanced accuracy equal to 61.74% when using the SVM linear classifier on the original dataset. The other classifiers resulted in poorer performance (less than 60%) when using the Multisurf* in all datasets.

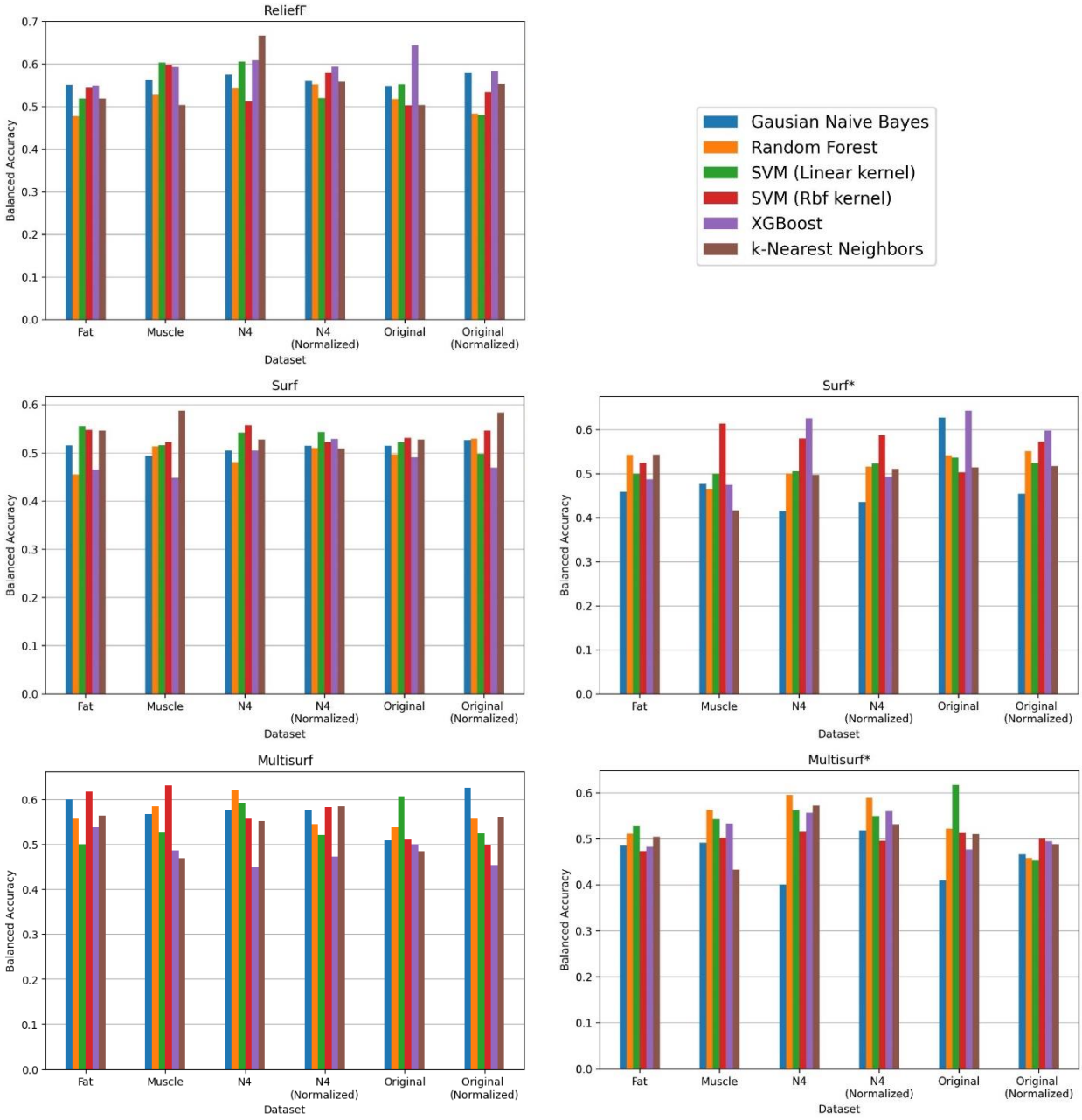

Figure 14. Balanced Accuracy for Relief-based algorithms per classifier and dataset

The pipelines of feature selection methods and classifiers that resulted in the optimal performance for each dataset are presented in Table 6.

For the original dataset, the optimal pipeline consists of Pearson correlation coefficient-based feature selection with threshold equal to 0.85, with 175 features (Appendix A), followed by the XGB classifier, achieving balanced accuracy of 70%. For the z-score normalised version of the original dataset, the Pearson-based selection method with 143 selected features (threshold equal to 0.85) and the SVM-Linear classifier were identified as optimal pipeline, achieving a balanced accuracy of 63.11%. The optimal pipeline for the N4 filtered version of the dataset consists of the ReliefF method using 68 features coupled with the kNN classifier,

achieving balanced accuracy equal to 66.70%. For both N4 normalised and fat-based normalised, the optimal pipeline includes the mRMR selection method with 23 selected features and the Random Forest classifier. In these two datasets, the optimal pipelines achieved balanced accuracy of 62.16%. Finally, for the muscle-based normalised dataset, the optimal pipeline includes the Multisurf selection method with 18 selected features and the SVM-RBF algorithm achieving a balanced accuracy equal to 63.09%. The N4 normalized and the fat-based normalized datasets achieved the lower balanced accuracy among the examined datasets, indicating weaker ability to identify the prostate cancer aggressiveness. In contrast, the optimal pipeline for the prediction of PCa aggressiveness consists of the univariate feature selection method with threshold equal to 0.85 and the XGB classifier on the original dataset.

Table 6. Optimal combination of feature selection method and classifier per dataset. The number of selected features along with the balanced accuracy are presented in the last two columns.

Dataset	Feature Selection Method	Classifier	# of Selected Features	Balanced Accuracy (%)
Original	Pearson/ Spearman (Thres=0.85)	XGB	175	70.00
Original with normalization	Pearson/ Spearman (Thres=0.7)	SVM (Linear)	143	63.11
N4 filtered	ReliefF	kNN	68	66.70
N4 filtered with normalization	mRMR	Random Forest	23	62.16
Fat-based normalization	mRMR	Random Forest	23	62.16
Muscle-based normalization	Multisurf	SVM (RBF)	18	63.09

4.2 Results of hybrid feature selection methods

After the completion of the analysis and the extraction of the optimal pipeline for each dataset, hybrid feature selection methods were also evaluated. The Pearson and the Spearman univariate methods were used as the first feature selection method to filter and exclude the highly correlated features. In the first experiment, this filtering was followed by the optimal pipeline derived from the initial analysis in order to predict the tumor's aggressiveness. However, in the original dataset and its normalized version, the optimal feature selection methods were also univariate methods. Thus, the second-best performing pipeline was used for the hybrid methods analysis. For the original dataset, the second-best performing pipeline consisted of the CMIM with 73 features coupled with the XGB classifier. For the normalized version, the pipeline includes the Multisurf with 13 features and the GNB classifier. These pipelines achieved balanced accuracy equal to 67.18% and 62.50% in the initial analysis, respectively. Furthermore, in a second experiment, the same optimal thresholds for the univariate feature selection methods were used, while GSCV was performed to identify the optimal number of features that the second feature selection method should select.

Table 7 presents the results per dataset for the two experiments using hybrid feature selection methods. Both experiments demonstrate decreased balanced accuracy score. On the original dataset, the pipelines achieved balanced accuracy larger than 60% in both experiments. However, all the pipelines demonstrated balanced accuracy in the range of 49.22% to 59.74% in the rest datasets. In addition, no significant improvement in the performance is observed while using GSCV for finding the optimal number of selected features for the second feature selection method.

Table 7. Balanced accuracy for hybrid feature selection methods per dataset. Optimal threshold is selected for the Pearson correlation coefficient-based feature selection method for each dataset based on the previous analysis.

Dataset	Balanced Accuracy	
	Optimal Threshold	Optimal Threshold and Feature Number
Original	61.81%	60.77%
Original with normalization	54.93%	57.40%
N4 filtered	59.61%	52.98%
N4 filtered with normalization	58.57%	49.22%
Fat-based normalization	57.27%	54.28%
Muscle-based normalization	58.44%	59.74%

Chapter 5: Discussion

In this study, various radiomic pipelines are developed to classify the aggressiveness of clinically significant prostate cancer. The main objective is to determine whether the cancer cells grow slowly or quickly enough for more efficient treatment planning. Image preprocessing techniques were applied to images to generate several filtered and normalised versions of the original PI-CAI dataset. In addition, 468 combinations of image preprocessing, feature selection methods and classifiers were employed to identify the optimal pipeline for detecting PCa aggressiveness. All pipelines were evaluated on a holdout test set. The analysis was extremely time-consuming, as the training and the testing of all possible pipelines in all datasets lasted for approximately 6 weeks.

After thoroughly examining the predictive power of all pipelines, the use of the original unnormalized dataset and the Pearson correlation coefficient-based feature selection method (threshold = 0.85) coupled with XGB classifier achieved a balanced accuracy of 70%. This pipeline achieved the highest performance among all the examined pipelines in detecting PCa aggressiveness. Pearson coefficient-based selection reduces the number of radiomic features from 1132 to 175 feeding the XGB classifier with the most informative features. Pearson and Spearman univariate filtering methods demonstrated the same performance using the same threshold and selecting the same features on both the normalized and unnormalized original datasets, implying the linear correlations between the radiomic features. The CMIM method with XGB achieved 67.18% balanced accuracy on the original dataset without any preprocessing, which is the second-best performing pipeline among all pipelines and datasets.

However, the third-best performance overall is achieved using the ReliefF algorithm for feature selection and the k-Nearest Neighbors classification algorithm on the N4 filtered dataset achieving a balanced accuracy equal to 66.70%. The values of the evaluation metrics in the N4 filtered dataset analysis are similar to their values in the original dataset. Hence, the bias field correction method does not negatively affect the ability of the models to detect the grade of the malignancy. Despite the promising results of the N4 filtering image preprocessing method, the normalized N4 filtered and the fat-based normalization methods demonstrated the lowest balanced accuracy of 62.16% among the optimal pipelines, using the mRMR and the Random Forest classifier. The z-score normalised version of the original dataset demonstrated balanced accuracy equal to 63.11% using the Pearson feature selection method (threshold = 0.7) and the SVM classifier with linear kernel. The use of a lower threshold implies the exclusion of a larger proportion of correlated features, indicating that the classifier requires features with low correlation between each other to make the optimal decision. However, none of the pipelines in this dataset outperformed the optimal pipeline trained in the original set. The second-best performing pipeline for this specific dataset consists of the Multisurf selection algorithm and the Naïve Bayes classifier achieving a balanced accuracy of 62.50%. Furthermore, the muscle-based normalized dataset achieved

balanced accuracy of 63.09%, which is close to the performance of the normalized original dataset. Thus, only the N4 filtering method does not degrade the model's performance.

The preprocessing techniques do not improve the performance of the classifiers in predicting the tumor's aggressiveness. Significant differences in the tumor's phenotype may be distorted by the preprocessing techniques affecting the radiomic features' values and thus degrading their ability to characterize the tumor. Although preprocessing techniques alter the pixel values for enhancing the visual interpretation of the images, in this context they seem to hamper the model's performance. Furthermore, none of the feature selection methods efficiently work for all classifiers. Accordingly, none of the classifiers efficiently work with all feature selection methods. Thus, the feature selection method and the classifier used for predicting the tumor's aggressiveness should be carefully selected. However, the XGB classifier is the only classifier that achieved a good performance (balanced accuracy larger than 60%) using all the feature selection methods except from the RBAs. The selected pipeline also depends on the dataset (preprocessed or not) in which to be applied as the signal intensities in the image are affected by the preprocessing technique. All classifiers, except for the kNN and the SVM rbf kernel, had optimal prediction results when using the original unnormalized data. The kNN and the SVM rbf kernel achieved their highest performance, but lower than the other classifiers in the original dataset, in the N4 filtered and the muscle-based normalized datasets, respectively. According to the results of the current study, the selection of the Pearson Correlation test with threshold equal to 0.85 and the use of the XGB classifier using the original dataset without any preprocessing are recommended for the challenging task of identifying the tumor's aggressiveness.

The use of the hybrid feature selection methods resulted in lower prediction performance. Pearson correlation coefficient feature selection method was used as the first filtering method of the pipeline, using the optimal thresholds that were extracted for each dataset from the initial analysis. The derived optimal feature method for each dataset was used as the second feature selection method coupled with the corresponding classifier for identifying the most informative radiomic features. The overall performance decreased in all datasets when executing the hybrid feature selection methods. Original dataset still achieves the best performance with a balanced accuracy of 61.81% followed by the N4 filtered dataset with a balanced accuracy of 59.61%. The normalized original, the normalized N4, the fat, and the muscle based normalized datasets achieved a balanced accuracy of 54.93%, 58.57%, 57.27%, 58.44%, respectively. This performance decrease occurs as the second feature selection methods were trained using different feature spaces. More precisely, the second feature selection methods were initially trained using the whole feature set (i.e., 1132 radiomics features). In the hybrid feature selection methods, the second feature selection method is fed with a significantly smaller feature space extracted after applying the first univariate method. Thus, the feature selection method when used as a second method may result in different selected radiomics compared to using it alone, as a different initial feature space is fed into the method.

Finally, the results are similar in the second experiment, where the second feature selection methods were trained using grid search 3-fold cross validation. Specifically, an increase in the balanced accuracy is observed when using the normalized original and muscle-based normalized datasets to 57.40% and 59.74%, respectively. In the rest datasets, the performance is decreased. The decrease in the performance may be due to application of the first selection method. More specifically, the first feature selection method eliminated a significant number of the original 1132 radiomic features, which may impact the selection capacity of the second feature selection methods. Furthermore, while applying GSCV, the optimal number of the selected features is tested on a small validation set and thus the optimal number may be different than the initially extracted one, which is used in the first experiment of the hybrid feature selection methods.

Overall, the limited performance of the hybrid feature selection methods may be due to the lack of optimization of the overall pipeline, instead of the second method only. The identification of the optimal threshold for the univariate method, which was used as a first step of feature filtering, was not investigated. The derived optimal threshold from the initial analysis was used in order to directly assess its effect on a hybrid feature selection method.

The optimal pipeline identified in the current study achieved a balanced accuracy of 70% and AUCROC of 0.70 in predicting the challenging task of prostate cancer aggressiveness. A similar study conducted by Sun et al. [43] investigated various feature selection methods and classifiers on glioma grading. The authors focused on a different pathology and anatomic region than in our study. However, they drawn the same conclusion that the predictive performance is affected by both the feature selection method and the classifier. Several other studies investigated the ability of radiomics and machine learning models to identify the clinically significant PCa from the non-clinically significant [42], [47]. Our aim was to identify the low aggressive tumor from the high aggressive tumor based on the cells growth. Thus, the results from these studies cannot be compared with our results, since the clinical and research question is different. Furthermore, Chaddad et al. [39] investigated a similar clinical problem to our study, predicting the grade group of the tumor based on the Gleason score. They achieved higher performance (AUC larger than 70%), but they used radiomics extracted from two modalities (i.e., T2w and ADC) and a significantly smaller dataset (i.e., 99 patients). Thus, the results cannot be directly compared as radiomics from different modalities were used. Moreover, Bertelli et al. [46] investigated the same clinical question to our study for predicting the prostate cancer aggressiveness based on same classification of the ISUP score. A slightly better performance was achieved having an AUC of 0.75. However, a significantly smaller dataset (i.e., 85 patients) was used than in our study (i.e., 220 patients). Thus, the robustness of these results are restricted. However, we achieved a classification performance close to the state-of-the-art performance for predicting the tumor's aggressiveness using a larger dataset than in the most studies. Hence, our study shows promising results in the challenging task of predicting the prostate cancer aggressiveness.

This study has several limitations. First, multiple iterations of the data splitting on train and test set should be performed in order to produce more robust results. Even though the data were split in a stratified manner and the train and test indices were the same for each dataset, models should be trained and tested for a predefined number of iterations in order to produce generalizable results. Second, the original dataset size is limited. More samples are required in order to build robust and well-validated datasets. Furthermore, the whole analysis of identifying the optimal pipeline for each dataset is an extremely time-consuming task, restricting the number of methods and classifiers to be examined. Moreover, multiple experiments were difficult to be performed due to the enormous execution time, making it difficult to reproduce the analysis. However, the preliminary results obtained from the current study when using the original dataset are promising for the classification of the prostate cancer aggressiveness.

Chapter 6: Conclusion

Classifying the aggressiveness of clinically significant prostate cancer is a challenging task. In this study, 468 combinations of different image preprocessing techniques, feature selection methods and classifiers were used to identify the optimal pipeline. Biologically motivated image preprocessing techniques and N4 filtering were used for enhancing the quality of the images. Results suggest that none of the image preprocessing steps could improve the performance of the classifiers in the specific task. Original unnormalized data demonstrated the best performance with a balanced accuracy of 70% (AUC: 0.70) using univariate filtering (Pearson and Spearman) feature selection methods and the XGB classifier. The normalized datasets achieved balanced accuracy in the range of 60% to 65% (AUC: 0.62 - 0.66), while the best performance among the preprocessed datasets achieved when using N4 filtered unnormalized dataset. Univariate filtering coupled with XGB and SVM with linear kernel achieved the optimal performance in the original and original normalized, respectively. ReliefF, Multisurf and mRMR coupled with KNN, SVM with Gaussian kernel and Random Forest, respectively, demonstrated optimal performance in the rest datasets. The optimal number of selected features varies from 18 to 175. Pearson and Spearman methods revealed a large number of selected features. Multisurf extracted a minimal feature set of 18 features, achieving balanced accuracy of 63.10%. Using hybrid feature selection methods, surprisingly, do not enhance the performance of the models. In contrast, the results retain a balanced accuracy in the range of 54.93% to 61.81% (AUC: 0.54 - 0.61). Optimizing the second feature selection method worsen even more the overall performance due to the restricted number of features.

Larger datasets could be used to validate the results of the current study. Multiple iterations could be used in order to increase the robustness of the results. Executing multiple runs and averaging the metrics will provide a clearer insight on the ability of the methods proposed to predict the tumor's aggressiveness. However, use of multiple runs requires increased computational resources and is extremely time-consuming; thus, it was not feasible to be performed in the current study. Furthermore, mpMRIs (T2W, DWI and ADC) available in the dataset would be used in order to assess whether the combined information obtained from all modalities enhance the performance of the models. Moreover, further feature selection methods and classifiers could be used to assess their performance. Tuning the hyperparameters of classifiers and optimizing the Pearson coefficient's threshold for hybrid feature selection are also recommended. Classifier stacking is another technique used for alleviating the poor performance of some classifiers, while utilizing the predictive power of the more powerful ones. Moreover, optimization of the dynamic pipeline for enabling parallel and distributed execution will be developed in future work in order to reduce the computational time. Furthermore, the use of a combination of both hand-crafted and deep-radiomics will be used in a future work to assess their joint predictive power in identifying the prostate cancer aggressiveness.

References

[1] C. V Berenguer, F. Pereira, J. S. Câmara, and J. A. M. Pereira, "Underlying Features of Prostate Cancer—Statistics, Risk Factors, and Emerging Methods for Its Diagnosis," *Curr. Oncol.*, vol. 30, no. 2, pp. 2300–2321, 2023, doi: 10.3390/curroncol30020178.

[2] M. B. B. Culp, I. Soerjomataram, J. A. Efstathiou, F. Bray, and A. Jemal, "Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates," *European Urology*, vol. 77, no. 1. Elsevier, pp. 38–52, 2020, doi: 10.1016/j.eururo.2019.08.005.

[3] M. J. Messina, "Emerging evidence on the role of soy in reducing prostate cancer risk.," *Nutr. Rev.*, vol. 61, no. 4, pp. 117–131, 2003, doi: 10.1301/nr.2003.apr.117-131.

[4] O. Kucuk *et al.*, "Phase II randomized clinical trial of lycopene supplementation before radical prostatectomy.," *Cancer Epidemiol. Biomarkers Prev.*, vol. 10, no. 8, pp. 861–868, 2001.

[5] N. Kurahashi *et al.*, "Dairy Product, Saturated Fatty Acid, and Calcium Intake and Prostate Cancer in a Prospective Cohort of Japanese Men," *Cancer Epidemiol. Biomarkers Prev.*, vol. 17, no. 4, pp. 930–937, 2008, doi: 10.1158/1055-9965.EPI-07-2681.

[6] J. E. McDunn *et al.*, "Metabolomic signatures of aggressive prostate cancer," *Prostate*, vol. 73, no. 14, pp. 1547–1560, 2013, doi: 10.1002/pros.22704.

[7] S. Salciccia *et al.*, "Biomarkers in prostate cancer diagnosis: From current knowledge to the role of metabolomics and exosomes," *Int. J. Mol. Sci.*, vol. 22, no. 9, 2021, doi: 10.3390/ijms22094367.

[8] N. J. Tustison *et al.*, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

[9] P. Dhal and C. Azad, *A comprehensive survey on feature selection in the various fields of machine learning*, vol. 52, no. 4. Applied Intelligence, 2022.

[10] M. C. Benson, I. S. Whang, C. A. Olsson, D. J. McMahon, and W. H. Cooner, "The use of prostate specific antigen density to enhance the predictive value of intermediate levels of serum prostate specific antigen.," *J. Urol.*, vol. 147, no. 3 Pt 2, pp. 817–821, 1992, doi: 10.1016/s0022-5347(17)37394-9.

[11] M. C. Benson *et al.*, "Prostate specific antigen density: a means of distinguishing benign prostatic hypertrophy and prostate cancer.," *J. Urol.*, vol. 147, no. 3 Pt 2, pp. 815–816, 1992, doi: 10.1016/s0022-5347(17)37393-7.

[12] E. Seaman, M. Whang, C. A. Olsson, A. Katz, W. H. Cooner, and M. C. Benson, "PSA density (PSAD). Role in patient evaluation and management.," *Urol. Clin. North Am.*, vol. 20, no. 4, pp. 653–663, 1993.

[13] H. B. Carter *et al.*, "Longitudinal evaluation of prostate-specific antigen levels in men with and without prostate disease.," *JAMA*, vol. 267, no. 16, pp. 2215–2220, 1992.

[14] H. P. Schmid, J. E. McNeal, and T. A. Stamey, "Observations on the doubling time of prostate cancer. The use of serial prostate-specific antigen in patients with untreated disease as a measure of increasing cancer volume.," *Cancer*, vol. 71, no. 6, pp. 2031–2040, 1993, doi: 10.1002/1097-0142(19930315)71:6<2031::aid-cncr2820710618>3.0.co;2-q.

[15] R. Raaijmakers *et al.*, “Prostate-specific antigen change in the European Randomized Study of Screening for Prostate Cancer, section Rotterdam.,” *Urology*, vol. 63, no. 2, pp. 316–320, 2004, doi: 10.1016/j.urology.2003.09.028.

[16] J. E. Oesterling *et al.*, “Serum prostate-specific antigen in a community-based population of healthy men. Establishment of age-specific reference ranges.,” *JAMA*, vol. 270, no. 7, pp. 860–864, 1993.

[17] J. S. Horoszewicz, E. Kawinski, and G. P. Murphy, “Monoclonal antibodies to a new antigenic marker in epithelial prostatic cells and serum of prostatic cancer patients.,” *Anticancer Res.*, vol. 7, no. 5B, pp. 927–935, 1987.

[18] R. Dhir *et al.*, “Early identification of individuals with prostate cancer in negative biopsies.,” *J. Urol.*, vol. 171, no. 4, pp. 1419–1423, 2004, doi: 10.1097/01.ju.0000116545.94813.27.

[19] H. Uetsuki, H. Tsunemori, R. Taoka, R. Haba, M. Ishikawa, and Y. Kakehi, “Expression of a novel biomarker, EPCA, in adenocarcinomas and precancerous lesions in the prostate.,” *J. Urol.*, vol. 174, no. 2, pp. 514–518, 2005, doi: 10.1097/01.ju.0000165154.41159.b1.

[20] T. J. Bradford, S. A. Tomlins, X. Wang, and A. M. Chinnaiyan, “Molecular markers of prostate cancer,” *Urol. Oncol. Semin. Orig. Investig.*, vol. 24, no. 6, pp. 538–551, 2006, doi: 10.1016/j.urolonc.2006.07.004.

[21] A. D. Choudhury *et al.*, “The role of genetic markers in the management of prostate cancer,” *Eur. Urol.*, vol. 62, no. 4, pp. 577–587, 2012, doi: 10.1016/j.eururo.2012.05.054.

[22] J. O. Barentsz *et al.*, “ESUR prostate MR guidelines 2012.,” *Eur. Radiol.*, vol. 22, no. 4, pp. 746–757, 2012, doi: 10.1007/s00330-011-2377-y.

[23] “PI-RADS™ Prostate Imaging and Reporting and Data System: 2015, version 2,” *American College of Radiology*. 2015.

[24] J. O. Barentsz *et al.*, “Synopsis of the PI-RADS v2 Guidelines for Multiparametric Prostate Magnetic Resonance Imaging and Recommendations for Use.,” *European urology*, vol. 69, no. 1. pp. 41–49, 2016, doi: 10.1016/j.eururo.2015.08.038.

[25] J. I. Epstein *et al.*, “The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System,” *Am. J. Surg. Pathol.*, vol. 40, no. 2, 2016.

[26] L. Egevad, B. Delahunt, J. R. Srigley, and H. Samaratunga, “International Society of Urological Pathology (ISUP) grading of prostate cancer – An ISUP consensus on contemporary grading,” *Apmis*, vol. 124, no. 6, pp. 433–435, 2016, doi: 10.1111/apm.12533.

[27] S. H. Hsu *et al.*, “Quantitative characterizations of ultrashort echo (UTE) images for supporting air-bone separation in the head,” *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2869–2880, 2015, doi: 10.1088/0031-9155/60/7/2869.

[28] L. Fang and X. Wang, “Brain tumor segmentation based on the dual-path network of multi-modal MRI images,” *Pattern Recognit.*, vol. 124, 2022, doi: 10.1016/j.patcog.2021.108434.

[29] S. Saman and S. J. Narayanan, “Active contour model driven by optimized energy functionals for MR brain tumor segmentation with intensity inhomogeneity correction,” *Multimed. Tools Appl.*, vol. 80, no. 14, pp. 21925–21954, 2021, doi: 10.1007/s11042-021-10738-x.

[30] M. Wang, J. Yang, Y. Chen, and H. Wang, “The multimodal brain tumor image segmentation based on convolutional neural networks,” *2017 2nd IEEE Int. Conf. Comput. Intell. Appl. ICCIA 2017*, vol. 2017-Janua, pp. 336–339, 2017, doi: 10.1109/CIAPP.2017.8167234.

[31] A. A. Nguyen *et al.*, “Post-Processing Bias Field Inhomogeneity Correction for Assessing Background Parenchymal Enhancement on Breast MRI as a Quantitative Marker of Treatment Response,” *Tomography*, vol. 8, no. 2, pp. 891–904, 2022, doi: https://doi.org/10.3390/tomography8020072.

[32] M. J. Saint Martin *et al.*, “A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study,” *Magn. Reson. Mater. Physics, Biol. Med.*, vol. 34, no. 3, pp. 355–366, 2021, doi: 10.1007/s10334-020-00892-y.

[33] A. Dovrou *et al.*, “A segmentation-based method improving the performance of N4 bias field correction on T2weighted MR imaging data of the prostate,” *Magn. Reson. Imaging*, vol. 101, no. March, pp. 1–12, 2023, doi: 10.1016/j.mri.2023.03.012.

[34] L. G. Nyúl and J. K. Udupa, “On standardizing the MR image intensity scale,” *Magn. Reson. Med.*, vol. 42, no. 6, pp. 1072–1081, 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[35] Z. Zhao, R. Anand, and M. Wang, “Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform,” *Proc. - 2019 IEEE Int. Conf. Data Sci. Adv. Anal. DSAA 2019*, pp. 442–452, 2019, doi: 10.1109/DSAA.2019.00059.

[36] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996, doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[37] M. Kursa, A. Jankowski, and W. Rudnicki, “Boruta - A System for Feature Selection,” *Fundam. Inform.*, vol. 101, pp. 271–285, 2010, doi: 10.3233/FI-2010-288.

[38] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *J. Biomed. Inform.*, vol. 85, no. January, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.

[39] A. Chaddad, T. Niazi, S. Probst, F. Bladou, M. Anidjar, and B. Bahoric, “Predicting gleason score of prostate cancer patients using radiomic analysis,” *Front. Oncol.*, vol. 8, no. 630, pp. 1–10, 2018, doi: 10.3389/fonc.2018.00630.

[40] M. Li *et al.*, “Radiomics prediction model for the improved diagnosis of clinically significant prostate cancer on biparametric MRI,” *Quant. Imaging Med. Surg.*, vol. 10, no. 2, pp. 368–379, 2020, doi: 10.21037/qims.2019.12.06.

[41] B. Liu *et al.*, “Prediction of prostate cancer aggressiveness with a combination of radiomics and machine learning-based analysis of dynamic contrast-enhanced MRI,” *Clin. Radiol.*, vol. 74, no. 11, pp. 896.e1--896.e8, 2019, doi: 10.1016/j.crad.2019.07.011.

[42] A. Rodrigues, J. Santinha, B. Galvão, C. Matos, F. M. Couto, and N. Papanikolaou, “Prediction of prostate cancer disease aggressiveness using Bi-parametric Mri radiomics,” *Cancers (Basel).*, vol. 13, no. 23, pp. 1–17, 2021, doi: 10.3390/cancers13236065.

[43] P. Sun, D. Wang, V. C. Mok, and L. Shi, “Comparison of Feature Selection Methods and Machine Learning Classifiers for Radiomics Analysis in Glioma Grading,” *IEEE Access*, vol. 7, pp. 102010–102020, 2019, doi: 10.1109/ACCESS.2019.2928975.

[44] A. Seetharaman *et al.*, “Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging,” *Med. Phys.*, vol. 48, no. 6, pp. 2960–2972, 2021, doi: 10.1002/mp.14855.

[45] I. Bhattacharya *et al.*, “Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework: CorrSigNIA: an MRI-pathology correlation and deep learning framework,” *Med. Image Anal.*, vol. 75, p. 102288, 2022, doi: 10.1016/j.media.2021.102288.

[46] E. Bertelli *et al.*, “Machine and Deep Learning Prediction Of Prostate Cancer Aggressiveness Using Multiparametric MRI,” *Front. Oncol.*, vol. 11, no. January, pp. 1–14, 2022, doi: 10.3389/fonc.2021.802964.

[47] J. M. Castillo T *et al.*, “Classification of Clinically Significant Prostate Cancer on Multi-Parametric MRI: A Validation Study Comparing Deep Learning and Radiomics.,” *Cancers (Basel).*, vol. 14, no. 1, 2021, doi: 10.3390/cancers14010012.

[48] A. Saha, J. Twilt, J. S. Bosma, M. Hosseinzadeh, and I. Slootweg, “Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI : The PI-CAI Challenge Imaging Data and Reference Standard,” 2022, doi: 10.5281/ZENODO.6667655.

[49] D. F. Gleason, G. T. Mellinger, and L. J. Ardving, “Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging,” *J. Urol.*, vol. 111, no. 1, pp. 58–64, 1974, doi: 10.1016/S0022-5347(17)59889-4.

[50] T. Y. Chan, A. W. Partin, P. C. Walsh, and J. I. Epstein, “Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy,” *Urology*, vol. 56, no. 5, pp. 823–827, 2000, doi: 10.1016/S0090-4295(00)00753-6.

[51] S. Song, Y. Zheng, and Y. He, “A review of Methods for Bias Correction in Medical Images,” *Biomed. Eng. Rev.*, vol. 1, no. 1, Sep. 2017, doi: 10.18103/BME.V3I1.1550.

[52] S. Gitto *et al.*, “Diffusion-weighted MRI radiomics of spine bone tumors: feature stability and machine learning-based classification performance,” *Radiol. Medica*, vol. 127, no. 5, pp. 518–525, 2022, doi: 10.1007/s11547-022-01468-7.

[53] E. Chang *et al.*, “Comparison of radiomic feature aggregation methods for patients with multiple tumors,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–7, 2021, doi: 10.1038/s41598-021-89114-6.

[54] J. Sled, A. Zijdenbos, and A. Evans, “A nonparametric method for automatic correction of intensity nonuniformity in MRI data,” *IEEE Trans Med Imaging*, vol. 17, no. 1, pp. 87–97, 1998, doi: 10.1109/42.668698. PMID: 9617910.

[55] R. T. Shinohara *et al.*, “Statistical normalization techniques for magnetic resonance imaging,” *NeuroImage Clin.*, vol. 6, pp. 9–19, 2014, doi: 10.1016/j.nicl.2014.08.008.

[56] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104--e107, 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[57] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994, doi: 10.1109/72.298224.

[58] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, “Conditional Likelihood Maximisation: A

Unifying Framework for Information Theoretic Feature Selection," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 27–66, 2012.

[59] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, 1992, pp. 249–256.

[60] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *AAAI*, 1992, vol. 2, pp. 129–134.

[61] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF BT - Machine Learning: ECML-94," 1994, pp. 171–182.

[62] M. ROBNIK SIKONJA MarkoRobnik and friuni-ljsi IGOR KONONENKO IgorKononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003.

[63] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *J. Biomed. Inform.*, vol. 85, pp. 168–188, 2018, doi: https://doi.org/10.1016/j.jbi.2018.07.015.

[64] C. S. Greene, N. M. Penrod, J. Kiralis, and J. H. Moore, "Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions," *BioData Min.*, vol. 2, no. 1, 2009, doi: 10.1186/1756-0381-2-5.

[65] C. S. Greene, D. S. Himmelstein, J. Kiralis, and J. H. Moore, "The Informative Extremes: Using Both Nearest and Farthest Individuals Can Improve Relief Algorithms in the Domain of Human Genetics BT - Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics," 2010, pp. 182–193.

[66] D. Granizo-Mackenzie and J. H. Moore, "Multiple Threshold Spatially Uniform ReliefF for the Genetic Analysis of Complex Human Diseases BT - Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics," 2013, pp. 1–10.

[67] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

Appendices

Appendix A – Optimal selected feature set from original dataset using Pearson correlation-base feature selection and XGB classifier.

"original_shape_Elongation"	"original_shape_Elongation"	"original_shape_Elongation"
"original_shape_Flatness"	"original_shape_Flatness"	"original_shape_Flatness"
"original_shape_MajorAxisLength"	"original_shape_MajorAxisLength"	"original_shape_MajorAxisLength"
"original_shape_Sphericity"	"original_shape_Sphericity"	"original_shape_Sphericity"
"original_firstorder_Minimum"	"original_firstorder_Minimum"	"original_firstorder_Minimum"
"original_glcm_Correlation"	"original_glcm_Correlation"	"original_glcm_Correlation"
"original_glcm_Imc1"	"original_glcm_Imc1"	"original_glcm_Imc1"
"original_glcm_Imc2"	"original_glcm_Imc2"	"original_glcm_Imc2"
"original_glszm_SmallAreaLowGrayLevelEmphasis"	"original_glszm_SmallAreaLowGrayLevelEmphasis"	"original_glszm_SmallAreaLowGrayLevelEmphasis"
"original_gldm_LargeDependenceHighGrayLevelEmphasis"	"original_gldm_LargeDependenceHighGrayLevelEmphasis"	"original_gldm_LargeDependenceHighGrayLevelEmphasis"
"original_gldm_SmallDependenceLowGrayLevelEmphasis"	"original_gldm_SmallDependenceLowGrayLevelEmphasis"	"original_gldm_SmallDependenceLowGrayLevelEmphasis"
"log-sigma-2-0-mm-3D_firstorder_Kurtosis"	"log-sigma-2-0-mm-3D_firstorder_Kurtosis"	"log-sigma-2-0-mm-3D_firstorder_Kurtosis"
"log-sigma-2-0-mm-3D_firstorder_Median"	"log-sigma-2-0-mm-3D_firstorder_Median"	"log-sigma-2-0-mm-3D_firstorder_Median"
"log-sigma-2-0-mm-3D_firstorder_Skewness"	"log-sigma-2-0-mm-3D_firstorder_Skewness"	"log-sigma-2-0-mm-3D_firstorder_Skewness"
"log-sigma-2-0-mm-3D_glcm_ClusterShade"	"log-sigma-2-0-mm-3D_glcm_ClusterShade"	"log-sigma-2-0-mm-3D_glcm_ClusterShade"
"log-sigma-2-0-mm-3D_glcm_Correlation"	"log-sigma-2-0-mm-3D_glcm_Correlation"	"log-sigma-2-0-mm-3D_glcm_Correlation"
"log-sigma-2-0-mm-3D_glcm_Imc1"	"log-sigma-2-0-mm-3D_glcm_Imc1"	"log-sigma-2-0-mm-3D_glcm_Imc1"
"log-sigma-2-0-mm-3D_glcm_Imc2"	"log-sigma-2-0-mm-3D_glcm_Imc2"	"log-sigma-2-0-mm-3D_glcm_Imc2"
"log-sigma-2-0-mm-3D_glcm_Idmn"	"log-sigma-2-0-mm-3D_glcm_Idmn"	"log-sigma-2-0-mm-3D_glcm_Idmn"
"log-sigma-2-0-mm-3D_glcm_InverseVariance"	"log-sigma-2-0-mm-3D_glcm_InverseVariance"	"log-sigma-2-0-mm-3D_glcm_InverseVariance"
"log-sigma-2-0-mm-3D_glrlm_ShortRunLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glrlm_ShortRunLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glrlm_ShortRunLowGrayLevelEmphasis"
"log-sigma-2-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"

"log-sigma-2-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-2-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-2-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"
"log-sigma-2-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-2-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-2-0-mm-3D_glszm_SmallAreaEmphasis"
"log-sigma-2-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"
"log-sigma-2-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"	"log-sigma-2-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"
"log-sigma-3-0-mm-3D_glcm_Correlation"	"log-sigma-3-0-mm-3D_glcm_Correlation"	"log-sigma-3-0-mm-3D_glcm_Correlation"
"log-sigma-3-0-mm-3D_glcm_Imc1"	"log-sigma-3-0-mm-3D_glcm_Imc1"	"log-sigma-3-0-mm-3D_glcm_Imc1"
"log-sigma-3-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-3-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-3-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"
"log-sigma-3-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-3-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-3-0-mm-3D_glszm_SmallAreaEmphasis"
"log-sigma-3-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"
"log-sigma-3-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_gldm_LargeDependenceHighGrayLevelEmphasis"
"log-sigma-3-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis"	"log-sigma-3-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis"
"log-sigma-4-0-mm-3D_firstorder_Skewness"	"log-sigma-4-0-mm-3D_firstorder_Skewness"	"log-sigma-4-0-mm-3D_firstorder_Skewness"
"log-sigma-4-0-mm-3D_glcm_Idmn"	"log-sigma-4-0-mm-3D_glcm_Idmn"	"log-sigma-4-0-mm-3D_glcm_Idmn"
"log-sigma-4-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-4-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"	"log-sigma-4-0-mm-3D_glszm_SizeZoneNonUniformityNormalized"
"log-sigma-4-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-4-0-mm-3D_glszm_SmallAreaEmphasis"	"log-sigma-4-0-mm-3D_glszm_SmallAreaEmphasis"
"log-sigma-4-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-4-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"	"log-sigma-4-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"
"log-sigma-5-0-mm-3D_firstorder_10Percentile"	"log-sigma-5-0-mm-3D_firstorder_10Percentile"	"log-sigma-5-0-mm-3D_firstorder_10Percentile"
"log-sigma-5-0-mm-3D_firstorder_Kurtosis"	"log-sigma-5-0-mm-3D_firstorder_Kurtosis"	"log-sigma-5-0-mm-3D_firstorder_Kurtosis"
"log-sigma-5-0-mm-3D_firstorder_Median"	"log-sigma-5-0-mm-3D_firstorder_Median"	"log-sigma-5-0-mm-3D_firstorder_Median"

"log-sigma-5-0-mm-3D_firstorder_Minimum"	"log-sigma-5-0-mm-3D_firstorder_Minimum"	"log-sigma-5-0-mm-3D_firstorder_Minimum"
"log-sigma-5-0-mm-3D_glcm_ClusterProminence"	"log-sigma-5-0-mm-3D_glcm_ClusterProminence"	"log-sigma-5-0-mm-3D_glcm_ClusterProminence"
"log-sigma-5-0-mm-3D_glcm_ClusterShade"	"log-sigma-5-0-mm-3D_glcm_ClusterShade"	"log-sigma-5-0-mm-3D_glcm_ClusterShade"
"log-sigma-5-0-mm-3D_glcm_Imc1"	"log-sigma-5-0-mm-3D_glcm_Imc1"	"log-sigma-5-0-mm-3D_glcm_Imc1"
"log-sigma-5-0-mm-3D_glcm_Idn"	"log-sigma-5-0-mm-3D_glcm_Idn"	"log-sigma-5-0-mm-3D_glcm_Idn"
"log-sigma-5-0-mm-3D_glcm_InverseVariance"	"log-sigma-5-0-mm-3D_glcm_InverseVariance"	"log-sigma-5-0-mm-3D_glcm_InverseVariance"
"log-sigma-5-0-mm-3D_glrlm_LongRunHighGrayLevelEmphasis"	"log-sigma-5-0-mm-3D_glrlm_LongRunHighGrayLevelEmphasis"	"log-sigma-5-0-mm-3D_glrlm_LongRunHighGrayLevelEmphasis"
"log-sigma-5-0-mm-3D_glszm_GrayLevelNonUniformity"	"log-sigma-5-0-mm-3D_glszm_GrayLevelNonUniformity"	"log-sigma-5-0-mm-3D_glszm_GrayLevelNonUniformity"
"log-sigma-5-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"	"log-sigma-5-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"	"log-sigma-5-0-mm-3D_glszm_LargeAreaLowGrayLevelEmphasis"