Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Master of Science and Technology in Reliable and Secure Systems	Autmn semester, 30 Open / Restricted access
Writer: Rihab Alzurkani	
Faculty supervisor: Álvaro Fernández Quílez	
Thesis title: Automated Computerized 2D and 3D Characterization of Tumors in Prostate Cancer	
Credits (ECTS): 30	
Key words: MRI, Radiomics, Shape Features Prostate Cancer, Machine learning	Pages: 64 Stavanger, 15/12/2023

University of Stavanger

Faculty of Science and Technology

Department of Electrical Engineering and Computer Science

Automated Computerized 2D and 3D Characterization of Tumors in Prostate Cancer

Master's Thesis in Computer Science

by

Rihab Alzurkani

Internal Supervisors

Álvaro Fernández Quílez

December 12, 2023

Abstract

The diagnostic pathway of Prostate cancer (PC) has been changed recently, and medical imaging has gained a significant role. Especially, Magnetic Resonance Imaging (MRI) has emerged as a pivotal tool in advancing cancer diagnosis. However, the interpretation of MRI data is encumbered by the intricate, time-demanding nature of the task and the inherent variability among readers, particularly radiologists. Considerably, achieving precise detection and thorough characterization of prostate tumours assumes paramount importance in facilitating subsequent procedures like guided biopsies. Convolutional neural networks (CNNs) have gained widespread traction in automating tasks encompassing classification and segmentation. Machine learning (ML) models can perform well as compared to deep neural networks, with, in some cases, the advantage of being easier to understand and interpret.

Given the importance of interpretability for clinicians to understand and interpret the features and decision-making process of the AI model, we select ML over deep learning (DL) for the early detection of prostate cancer from MRI images. The principal intent of this research is to serve as a starting point to the development of a computer-aided diagnosis system on ML. We employ prostate features instead of tumour features for an early detection of PCa. With this approach, we aim to avoid issues related to the difficulty of detecting and segmenting small tumour structures and their subsequent characterization problems. The main goal of the proposed model is to achieve a better strategy for detecting prostate cancer and accurately selecting the suitable decision for treatment procedures. The proposed model represents a significant leap forward in the realm of prostate cancer detection and diagnosis. The current strategies offer enhanced accuracy and reliability as compared to deep learning in identifying prostate cancer.

Our analysis indicates that logistic regression (LR) emerges as the optimal machine learning model for our objectives, as evidenced by its superior accuracy (0.88) and ROC-AUC score (0.93). These metrics underscore LR's reliability and interpretability, key factors in its effectiveness. This study proposes that our machine learning system could significantly improve the detection of prostate cancer, thereby offering healthcare professionals critical insights to inform personalized treatment strategies.

Acknowledgements

Firstly, I would like to thank and appreciate God, the Almighty for His grace and gift of life and sound health to be able to run through and complete my master's Program. Diligent efforts have been put into the completion of this thesis work. However, it would not have been possible without the support and guidance of my amiable and ever-supportive supervisor Prof. Alvaro Fernandez. I would like to extend my sincere thanks to him.

I also would like to thank Maher. You are well appreciated. I would like to dedicate this work to the loving memory of my dear father, Hashim, who recently left this world, and was a constant source of inspiration and encouragement throughout my academic journey. His wisdom, love, and unwavering support have left an indelible mark on my life and the completion of this thesis is a tribute to his memory. Though he is no longer with us, his spirit continues to guide and inspire me. I will always be grateful for the profound impact he had on my life.

Finally, I wish to express my gratitude towards my family for their cooperation, encouragement, endurance, perseverance, and support which has really helped in the completion of this thesis

Contents

List Of Figures

List of Tables

Chapter 1

1.1 Research and Motivation

The research is driven by the motivational development of an automated computerized system that employs improved techniques in image processing (currently 2D and will be for 3D) and analyzing the characteristics of the tumor in the prostate. The utilization of the state of art technology and algorithms to extract, analyze, and interpret critical information from MRI data, the main research objectives are summarized in the following:

1. Feature Extraction: extract the shape features to capture a wide range of quantitative data from the tumor region including texture attributes and intensity characteristics.
2. Classification: Utilize machine learning ML and artificial intelligence AI models to extract tumor characteristics and the difference between benign and malignant cases.
3. Gleason Score GL: classified based on the Gleason score which plays a crucial role in determining the aggressiveness of prostate cancer.
4. Tumor Segmentation: This is further work in this project to develop ML algorithms for accurate segmentation of tumors in 3D scans, to detect the precise delineation of tumor boundaries.

Many factors motivate the research in Automated Computerized 3D Characterization of tumors in Prostate Cancer. Firstly, prostate cancer is widespread, and the condition needs automated tumour characterization to raise the probability of early detection, correct diagnosis, and improved treatment decisions. Secondly, the 3D nature of MRI data, with rich in tumour details shows the complexity that can be addressed through automation to ensure consistent and efficient analysis. Thirdly, tumour automation is vulnerable to the custom of treatment plans, simply making the treatments fit the tumour condition will improve the cancer treatments for each patient. Lastly, the developments in artificial intelligence offer pricing in enhancement in tumour characterization.

1.2 Objectives:

The primary objective of this master's thesis, involves three key goals:

1. Extract Shape features (Pyradiomics Feature) from MRI Images:
   The first objective is to competently extract Pyradiomics features from magnetic resonance imaging (MRI) scans of patients, with an instance emphasis on shape-related features. These Pyradiomics features will act as essential data points for subsequent analysis.
2. Classify Patients:
   The second objective involves leveraging the extracted Pyradiomics features to acquire a robust classification model. This model will be instrumental in accurately classifying between patients with prostate cancer and those who are healthy. Reaching this classification task will contribute significantly to early and specific cancer detection.
3. Generate 3D Mesh Tumour:
   The third objective revolves around the generation of mesh structures indicating tumours. This task is imperative not only for the immediate aims of the study but also as a valuable foundation for forthcoming projects. The derived Mesh features will serve as an integral component in enhancing our understanding of tumour characteristics.

By accomplishing these objectives, this master's thesis aims to advance the field of prostate cancer diagnosis by automating the 3D characterization of tumours through state-of-the-art Pyradiomics feature extraction, patient classification, and the creation of tumour Mesh structures. These outcomes hold promise for improving the efficiency and accuracy of cancer diagnosis and may have broader implications for oncological research and patient care.

1.3 Related work

The keys studies are integrated to provide a comprehensive understanding of this field, Cuocolo et al. (2019) , highlighted the diagnostics potential of MRI-derived shape features which is the same in our objective" in this study, particularly in his work, "Clinically significant prostate cancer detection on MRI: A radiomic shape features "study published at European Journal of Radiology, he focused on extracting the surface to the volume ratio(SVAR), in identifying significant prostate cancer with Gleason score values for classification [1]. In Cuocolo work achieved highest values for accuracy and sensitivity with 95% confidence intervals.

In our study we adopted methodology like Shanker et al. (2023) for calculating accuracy and F1-score. Though, Shanker et al. analyze the morphological features related to lymphoma subtypes, specifically nuclear features, our approach diverged by concentrating on extraction shape features [2]. The studies in lymphoma classification the subtype with high accuracy more than 90%using deep learning during 2 to 4 subtypes included. In our study as comparable the model achieved 88% accuracy for Logistic Regression without resampling.

This aligns with the works of Ning et a., demonstrate the effectiveness of dual-tracer PET/CT scans in distinguishing between benign and malignant pulmonary using radiomics analysis or improved diagnosis accuracy [3].Extending this concept, Stoyanova et al. (2016) explored a radio genomics approach by combining radiomic data from mpMRI with genomic data [4], aiming to enhance the precession of prostate cancer treatment.

A related work in the field of vivo tumor characterization using machine learning has been conducted by Krajnc et al. (2022). Their study published in Frontiers in Oncology, provides insights into data preparation methods that enhance the accuracy of cancer characterization. This approach validated in glicoma, prostate, and diffuse large B-cell lymphoma cohort, using SMOTE. The random forest shows 0.79 accuracy, while we achieved in our work for same model 0.74. These findings demonstrate improvements in performance especially in Random Forest and Support Vector machine schemes [7].

In this thesis. Build on those findings, employing machine learning techniques with MREI-extracted shape features to refine prostate cancer diagnostics, emphasizing the importance of fewer, impactful features for clinical efficiency and accuracy.

Chapter 2

2.1 Medical Background

Cancer of the prostate is the highest rate of cancer among men in Norway. Annually, about 5000 people get this diagnosis. Norway, in 2022, prostate cancer was diagnosed in 5,474 men, translating to a rate of 183.6 cases for every 100,000 individuals according to the Cancer Register of Norway.

From 2018 to 2022, early, localized-stage detection accounted for 53.5% of these cases. Diagnoses during the regional stage made up 29.5%, and 8.9% of the men had advanced to distant metastasis when diagnosed. The stage of diagnosis remained undetermined for 8.1% of the cases.[8]

Prostate cancer predominantly impacts senior men, with its prevalence rising in tandem with the ageing population. This malignancy emerges from the ducts and glands of the prostate gland. In its early stages, the disease frequently exhibits no noticeable symptoms.

From the early 1990s through the mid-2000s, there was a notable surge in diagnosed prostate cancer cases across all age demographics, except for those aged 0-54 years. As shown in the figure shows a high rate of detection of cancer from 2020 to 2021 and less from 2021 to 2022.

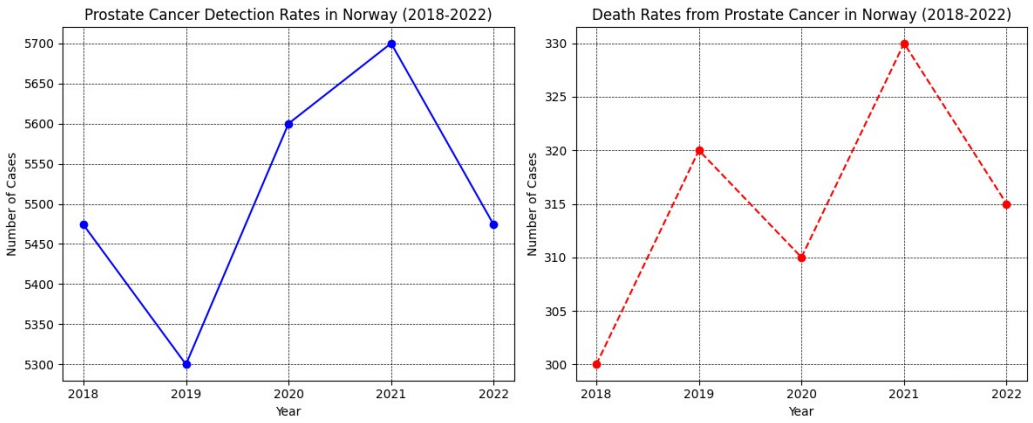


Figure 2.1 shows the cancer detection rates in Norway during the period 2018 until 2022 [8]

The exact origins of prostate cancer remain uncertain. Genetic factors are suspected to influence up to 10% of cases. A heightened risk exists for individuals with multiple family members (like brothers, fathers, or grandfathers) who have been diagnosed with or have battled the condition,

especially if these relatives were diagnosed before the age of 65-70. As the age is important factor besides another factors like shape features [9].

Figure 2. shows two-line graphs related to prostate cancer in Norway between 2018 and 2022. The left graph, with a blue line, represents detection rates with a peak in 2020. The right graph, with a red line, shows death rates with a peak in 2021. Both graphs show year-on-year fluctuations in the number of cases. Figure 3 shows trends in incidence (dark blue), mortality (light blue) and 5-year comparative survival (green) from prostate cancer in the interval 1965-2022. From figure 9.1-O in Cancer in Norway 2022.

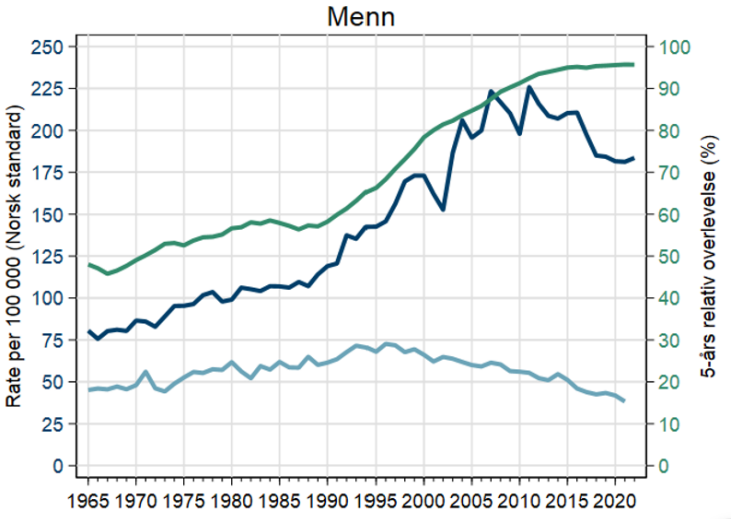


Figure 2.2 Rate of Cancer in Norway 2020 [8]

2.2The Prostate anatomy

The prostate is a walnut-sized gland located below the bladder in men and surrounds the urethra. It plays a vital role in male reproductive health, mainly responsible for producing seminal fluid that nourishes and carries sperm. Regarding cancer detection, the prostate becomes especially significant, as it can be the origin of malignant tumors. Clinicians frequently examine the prostate's size, shape, and texture when screening for potential cancer. Additionally, blood tests looking for specific prostate cancer markers can further enhance detection accuracy. Early identification of any aberrations in the prostate can be pivotal for the effective treatment and management of prostate cancer according to American Cancer Society, Cancer statistics center [12].

The male generative system, the prostate gland is made up of around 40-50 glands that encircle the urethra. These glands are classified into inner and outer clusters, parted by fibrous tissue. The main task of it is responsibility for creating seminal fluid, and the last is fundamental for nourishing and safeguarding sperm. This fluid remains liquid by the glycoprotein called prostate-specific antigen (PSA), produced by the prostate's epithelial cells. The higher PSA in the bloodstream may refer to

conditions such as prostate cancer, enlargement, or infection [12]. The prostate is anatomically separated into three distinct zones, each with its distinctive structure originating from the urethra. These are the peripheral zone (PZ), central zone (CZ), and transitional zone (TZ). In some contexts, the CZ and TZ together are referred to as the central gland as shown in figure 4 below:

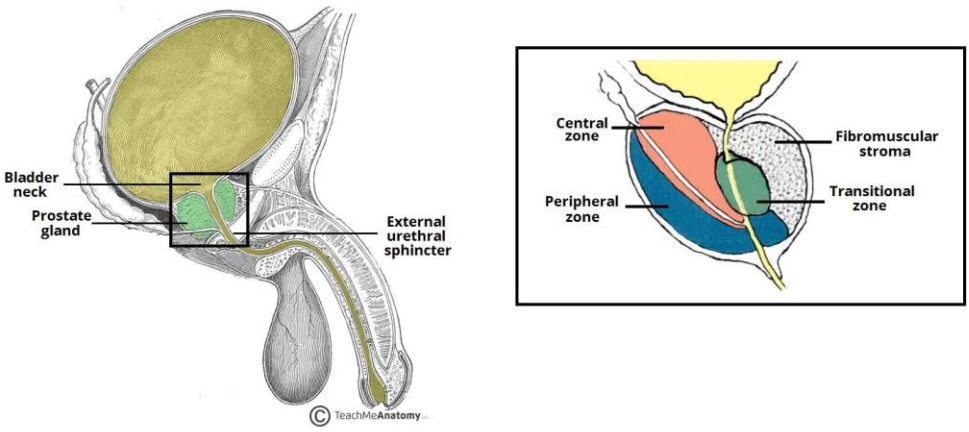


Figure 2.3 Diagram showing the zonal anatomy of prostate- transition zone (TZ), central zone (CZ), peripheral zone (PZ) and the anterior fibromuscular stroma (FMS) with respect to the prostatic urethra (PU) and the ejaculatory ducts (ED) [12].

The peripheral zone (PZ), forms up about 70% of the gland and is regularly someplace cancer may arise; the transitional zone (TZ), starts as 5-10% of the gland in younger men but can rise with age because of benign prostatic hyperplasia (BPH); and the central zone (CZ), creation up to 25% of the gland, mostly containing dense tissues and complex glands, and is least prone to diseases [13].

2.3AI and Prostate Cancer

With recent advances in AI, prostate cancer can be perfectly characterized and more optimally managed according to tumor biology, patient preferences, and survivorship goals. Advances in prostate cancer detection and management have markedly improved the precision of forecasting outcomes for those affected by the condition [13].

Better methods for categorizing risk, improved imaging technologies, and the strategic use of biomarkers have all contributed to a deeper understanding of the disease's likely progression in individuals [14][4].In the therapeutic landscape of prostate cancer, a widely recognized strategy involves the careful observation of the disease's evolution, with the aim of timely application of local treatments. Surgical and radiotherapeutic methods are continuously being enhanced, providing a sharper focus on minimizing potential negative side effects from treatments. Additionally, there's been

a significant improvement in life expectancy for men with advanced prostate cancer, with median survival times now reaching up to five years. This increase is largely due to the early introduction of drugs like docetaxel, abiraterone, enzalutamide, and a range of other novel agents [16].

2.4Standard detection and diagnosis of PC

The diagnosis of PCs can be complicated due to the potential side effects or limitations of the test. There is no single dedicated test that conclusively reveals it. As an alternative, a sequence of tests, each with characteristics, is employed when symptoms are established. These various techniques together lead to the diagnosis shown in Figure 2.4 stages of PC detection. The provided flow in the image typically goes from a Digital Rectal Exam (DRE) to a Prostate-Specific Antigen (PSA) test, followed by a biopsy, and then Magnetic Resonance Imaging (MRI). However, there can be scenarios where the order of a biopsy and MRI might be switched based on specific clinical situations or protocols.

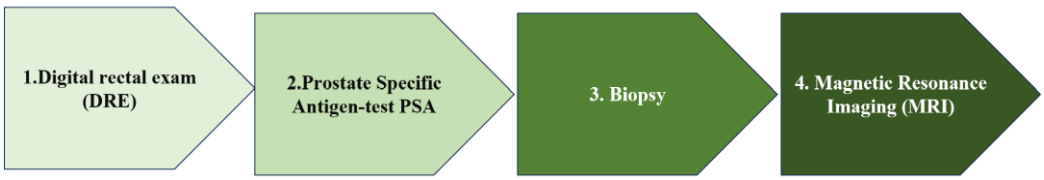


Figure 2.4 Different stages of PC detection

2.4.1 Digital rectal exam (DRE)

In case the patient shows signs of potential prostate cancer, the physician will inquire about symptoms and their duration, as well as any risk factors, including family history. A digital rectal exam (DRE) can be performed, where the doctor checks the prostate for abnormalities. This exam can indicate the cancer's extent. Based on the findings, further tests may be recommended. The results of the DRE and PSA tests cannot diagnose prostate cancer. The diagnosis is based on a tissue sample positive for cancer cells [9][11].

2.4.2 Prostate Specific Antigen-test PSA

The Antigen-test or (PSA) test is the blood examination to measure the level of PSA, a protein produced by prostate glands in the blood of the patient. While this test can indicate the possibility of PC, it is not definitive in diagnosing PSA in blood can be affected by different factors, making the test less than completely reliable for cancer detection. Consequently, there is no full consensus among the researchers regarding the efficiency of the PSA test in improving the detection of PCs. Ongoing research aims to refine and enhance the methods for detecting

and diagnosing this cancer. Attempting for more accurate and reliable techniques. This information is further detailed by the National Cancer Control Program (NCCP) prostate pathway Subgroup in their patient information guide, “Deciding about the Prostate Specific Antigen (PSA) Test”, provided by the Health Service Executive (HSE) health and social care [10][12].

2.4.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) stands as a pivotal non-invasive diagnostic tool, especially in the realm of prostate cancer detection. This technique, distinct for its avoidance of ionizing radiation, yields detailed cross-sectional images of the human body. For prostate cancer, MRI's exceptional ability to provide high-contrast images of soft tissues is invaluable. It leverages multiple contrast parameters and can capture images in various orientations, offering both 2D and 3D data. This versatility underlines MRI's integral role in diagnosing prostate cancer. Since its introduction, MRI has witnessed widespread adoption. By 2008, it was estimated that between 20,000 and 24,000 MRI systems were operational worldwide, facilitating 60 to 80 million MRI examinations each year. These statistics are elaborated in [17] study on MRI methodologies.

2.4.4 Biopsy

However, the PSA test can sometimes give misleading results, as high PSA levels can also be caused by benign conditions, medical procedures, or infections, not just cancer. It's also important to note that many prostate cancers are found through screening, which may progress too slowly to pose a significant threat to life, leading to potential overtreatment. The biopsy procedure is generally safe but can have side effects. The most common is hematuria, or blood in the urine, occurring in about 31% of cases. More serious, but less common, complications include infection (0.9%) and, in rare instances, death (0.2%). These insights and statistics are detailed in the study “Recommendations on screening for prostate cancer with the prostate-specific antigen test[18]. To conclude the best treatment methodology and estimate the effect for prostate cancer, it's important to evaluate two key factors:

- **Microscopic examination of the tissue sample:** This examination implies analyzing a small piece of tissue from the suspected area under a microscope. The sample examination classifies the cellular characteristics, checks if cancer cells are present, and decides the type and aggressiveness of the cancer.
- **Clinical staging of the index tumour and its potential metastases:** Assessing the tumor size and detecting whether the cancer has been distributed to nearby lymph nodes or other parts of the body beyond the origin. This will provide insight into the best treatment approach and recommend predictions on patient prognosis [10].

2.5 Gleason score

The biopsy tissue undergoes microscopic analysis to receive a Gleason score, GS is a grading system used to assess the aggressiveness of prostate cancer based on the appearance of cancer cells in a prostate tissue sample obtained from a biopsy a prevalent method for classifying prostate cancer severity. It was developed by Dr. Donald Gleason in the 1960s and is an essential tool for diagnosing and staging prostate cancer[20].

The Gleason system, depicted in Figure 2.5 categorizes cancer cells based on their similarity to normal prostate cells, with grades ranging from 1 to 5. In grades 1 and 2, cancer cells are closely akin to normal cells. In contrast, higher grades show significant differences, with a higher grade indicating a potentially more aggressive tumor. A pathologist assigns two grades to a tumor: a **primary** grade for the dominant pattern and a **secondary** for the next prevalent one. These two grades are then combined to produce a Gleason score, which helps determine the cancer's aggressiveness.

1. Gleason Grade 1: The cancer cells closely resemble normal prostate tissue and are less aggressive.
2. Gleason Grade 2: The cancer cells still have a relatively well-formed glandular structure and are also less aggressive.
3. Gleason Grade 3: The cancer cells are more disorganized and have a higher chance of being aggressive.
4. Gleason Grade 4: The cancer cells have an even more abnormal appearance and are moderately aggressive.
5. Gleason Grade 5: The cancer cells are highly abnormal and indicate the most aggressive form of prostate cancer.

The pathologist identifies the two most prevalent patterns of cancer cells in the tissue sample, and these two Gleason grades are then added together to create the Gleason score. The final Gleason score is typically reported on a scale of 6 to 10. For example: If the predominant pattern is grade 3 and the secondary pattern is grade 4, the Gleason score is 3 + 4 = 7. If both patterns are the same, say grade 4, then the Gleason score is 4 + 4 = 8.

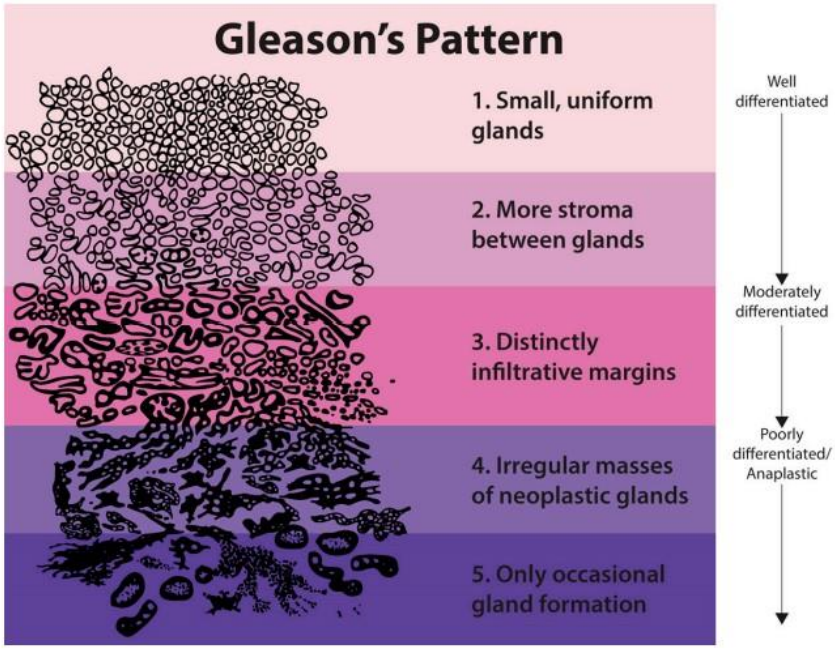


Figure 2.5 Gleason grading system depiction [19]

A lower Gleason score (6 or less) suggests that the cancer is less aggressive and may have a better prognosis. Gleason scores of 7 are considered intermediate-grade cancer, and Gleason scores of 8 to 10 are considered high-grade cancer, indicating a higher risk of aggressiveness and potential for spreading.

The Gleason score plays a crucial role in determining the stage of prostate cancer and helps guide treatment decisions. It provides valuable information to physicians about the likelihood of cancer spreading beyond the prostate and helps them formulate a personalized treatment plan for each patient. However, treatment decisions are not based solely on the Gleason score, and other factors, such as the cancer stage, PSA levels, and the patient's overall health, are also considered to create the most appropriate treatment approach [20].

Chapter 3 Technical Background

3.1 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) is a type of imaging scan, used in medical scanning procedures. MRI uses magnetic fields and radio waves to make pictures to take pictures of the body's interior. It scans image water, as all tissue contain different amounts of water which make them so efficient. This allows high-resolution pictures of different organs and tissues that are invisible to X-ray imaging [21].

The MRI originated from Nuclear Magnetic Resonance (NMR), which investigates the physical and chemical characteristics of atoms. In 1970 was first MRI was used on humans. Later it has many improvements while it provides low-resolution images. Over time various enhancements improve the image resolution and reduce the number of scan times. These improvements in magnet design, radio frequency, and applied algorithms. The first MRI was created in 1980 in Nottingham and Aberdeen and currently, it is widely used as an efficient tool for diagnosis in the medical field [22][23].

The fast growth in medical applications has been supplemented by various technology-related enhancements in MR imaging over the years. The noticeable improvement in terms of image clarity and diagnosis capabilities is with High field MRI (1.5 and above) in the 1990s. In the early nineteenth functional MRI (FMRI) allowed to map from the brain function by detecting changes in blood flow. During the 2000s Diffusion Tencer Imaging MRI (DTI) prominent imaging neural pathway, also diagnosis of stroke and stages of cancer besides Cardiac MRI aiding in the assessment of cardiac diseases [24].

During the 2010s, 3T and 7T Tesla(7T) provided more details for research purposes. Further changes to achieve better characteristics like Silent MRI in reducing the noise of scanning to achieve patient comfort. Despite the developments, there are challenges and limitations, MRI is still high cost with less accessibility in specific areas [25]. Throughout this upgrade, MRI has continuously progressed, MRI offering increasingly detailed views of the human body without ionizing radiation, presenting it as one of the most versatile and powerful diagnostic tools in medicine.

After this short introduction, MRI quickly emerged as a vital imaging method, and it continues to play an important role in health care. Its prominence can be attributed to unique qualities, especially with soft tissue contrast derived from various contrast parameters, the capacity to capture images from different angles, and the ability to handle both 2D and 3D data. MRI stands for non-invasive diagnostic imaging technique that produces detailed cross-sequential images of the human body, all without employing ionizing radiation.

3.2 Insights into Magnetic Resonance Imaging (MRI)

MRI machines work principle the same as electromagnets, the electrical current is passed through a large coil and generates an electromagnetic and the coil is cooled to an extremely low temperature using liquid helium (-273 °C), transitioning it into a superconducting state to eliminate the electrical resistance. This will generate the magnetic field used in mainstream MRI machines at 1.5 or 3.0 Tesla (T), a unit of magnetic field strength.

The notable drawback is superconducting magnets need bulky equipment and periodic replenishment of helium coolant. This will make the patient feel uncomfortable therefore open gantry MRI machine style. which placed a permanent magnate this provides more space for patients and more comfortable. However, there is a trade-off in magnetic field strength, as permanent magnets are limited to generating fields of a round 0.5 Tesla[26]. Magnetic resonance Imaging is non-intrusive, and it can provide functional information. In MRI, T1-T1-weighted images highlight fatty tissue by making it appear brighter, while T2-weighted images make water-rich areas stand out. those contrasts help to distinguish tissues differently and for abnormality detection.

3.3 Clinical Application of Magnetic Resonance Imaging of the Prostate

Clinical applications of Magnetic Resonance Imaging (MRI) for the prostate are significant and diverse [21][23]. MRI plays a crucial role in prostate cancer diagnosis, staging, and treatment planning. It is used to:

1. **Prostate Cancer Detection**: MRI is highly sensitive in detecting prostate cancer, especially in cases of clinically significant tumours. It helps identify suspicious lesions within the prostate gland.
2. **Localization**: MRI assists in precisely localizing the tumour within the prostate, helping guide biopsy procedures for accurate diagnosis.
3. **Staging**: MRI provides detailed information about the extent of cancer, including whether it has spread beyond the prostate gland (staging), aiding in treatment decisions.
4. **Active Surveillance**: For low-risk prostate cancer cases, MRI is used for active surveillance to monitor disease progression without immediate intervention.
5. **Treatment Planning**: MRI is essential for treatment planning, such as guiding targeted radiation therapy and assisting in surgical planning for prostatectomies.
6. **Assessing Treatment Response**: After treatment, MRI can assess the response of the tumor to therapies, helping to evaluate treatment effectiveness.
7. **Evaluating Recurrence**: MRI is valuable in detecting prostate cancer recurrence after initial treatment, assisting in early intervention.

8. **Assessing Pelvic Anatomy**: Beyond cancer, MRI can assess the overall pelvic anatomy, helping diagnose and manage other conditions affecting the prostate and surrounding structures.

3.3.1 T2 Weighted Images

T2 Weighted images is a basic pulse sequence in magnetic resonance (MR) imaging that despite the alteration in T2 time relaxation of various tissues. T2-weighted images have the best selection of diagnosis diseases because most tissues involved in a pathologic process have higher water content than normal and the liquid causes the affected regions to show brighter on T2-weighted images [26].

Diffusion-weighted imaging (DWI) is a specialized MRI technique that captures the movement of water molecules within tissues. It creates images based on the rate of water diffusion, where restricted diffusion appears bright and more free diffusion appears dark. DWI is widely used in medical imaging to diagnose and assess various conditions, including stroke, brain tumors, and abdominal abnormalities [27].

The apparent diffusion coefficient (ADC) map derived from DWI quantifies diffusion rates in tissues, aiding in the characterization of lesions. Additionally, diffusion tensor imaging (DTI), a more advanced form of DWI, offers insights into neural pathways in the brain. Overall, DWI is an essential tool in modern medicine, providing crucial information about tissue microstructure and pathology [27]. T2-weighted images are a fundamental component of multiparametric resonance imaging(mpMRI) because they produce high-resolution images that depict prostate anatomy. These types of images are important the reason belong of their ability to create contrast between different soft tissues in the body aiding in diagnosis. T2-weighted images highlight tissue variations in T2 relaxation times, with longer T2 times appearing bright and shorter T2 times appearing dark. This contrast is valuable for visualizing soft tissues and identifying abnormalities like edema, inflammation, tumors, cysts, and lesions. T2-weighted images are widely used in medical diagnosis across different body regions, including the brain, spine, abdomen, pelvis, and extremities. They also help differentiate tissues within the brain and are generated using specific MRI pulse sequences with longer echo times (TE). In summary, T2-weighted images are crucial in MRI for their ability to provide essential contrast information in diagnosing a range of medical conditions [29].

3.4Radiomics

Radiomics is a valuable tool for extracting clinical information from radiological imaging. There are various studies in medical imaging with radiomics, the basis is to extract from medical images

quantitative and recordable data. The images are difficult to recognize or diagnose by human eyes. Throughput feature extraction predicates the patient case with a large training cohort to get stable information from image characteristics and the disease stage. There are different scenarios for feature extraction from metabolic imaging like PET and SPECT [30].

While radiomics is a powerful tool, its data-centric approaches provide no intuition into the underlying biological means of taking the observed correlation [4]. The significant role of the radiomic analysis features and parameters in characterizing the tumor response is still not achieved due to the limitation of this technique. With current developments in machine learning approaches, this field has improved into High-throughput agnostic analysis [31].

3.5 Radiomic Feature Types and extraction

Radiomic features can be summarized into two categories: statistically histogram-based and texture-based, as described in [31]. Additionally, our specific interest lies in transformed-based and shape-based features. In this study, we aim to focus on the region of interest (ROI) for extracting both 2D and 3D features to enhance readability, following the approach outlined by [1][3][2]. Statistical structures are applied to unmodified or discretized grayscale level intensities. The aggregation of gray level and feature values is not minimized to improve robustness and reproducibility, following the methodology proposed by [31].

- First order statistics
- Shape-based(3D)
- Shape-based(2D)
- Gray Level Co-occurrence Matrix
- Gray Level Run Length matrix
- Gray Level run length Matrix
- Gray Level Size Zone Matrix
- Neighboring Gray Tone Difference matrix
- Gray level dependence Matrix

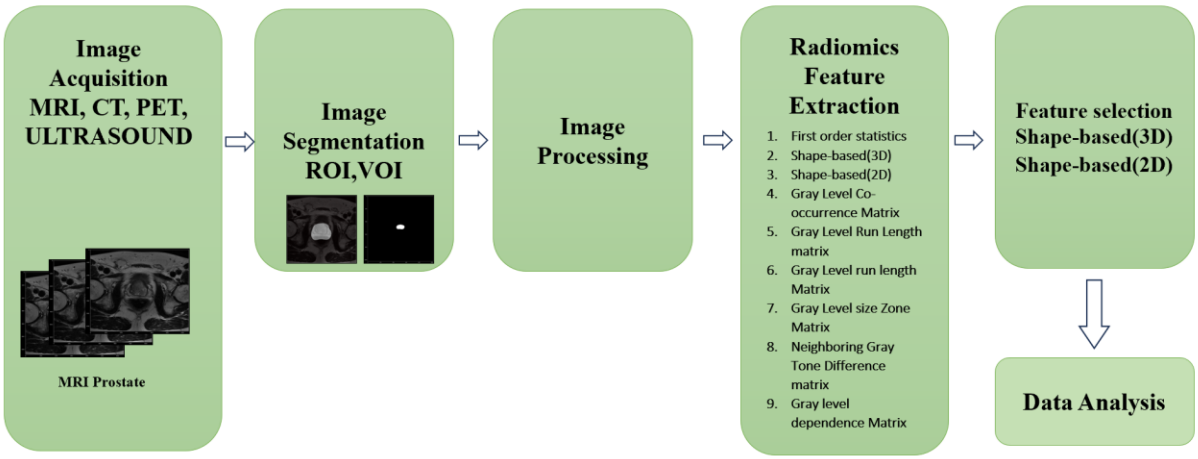


Figure 3.1 Extracting Radiomics

The implementation of Radiomics approach pipeline consists of several steps as shown in Figure3.1 Generally beginning with Image acquisition, preprocessing-image, defining ROI region of interest or as it is known image segmentation, and for further classification applying machine learning algorithms for disease classification, feature detection, and variable response prediction [32].

Step1: Image acquisition: This process of capturing medical images utilizes various modalities such as MRI, Magnetic Resonance Imaging), CT (computed tomography, PET (positron emission Tomography), or Ultrasound. The primary goal is to create a digital image of patient anatomy or pathology [21][32].

Step2: Image segmentation: Region of interest ROI for 2D or volume of interest VOI for the 3D approaches is a crucial step for images. Various software solutions either commercial or open source like the 3D slicer we used for segmentation, MITK, ITK-SNAP, Me ViSLAB, lifEx, or ImageJ.

Step3: Image processing Image processing in radiomics requires steps to be prepared for feature extraction and those steps are crucial to ensure the accuracy and reliability of feature extraction. It includes image loading and resampling to ensure all images have voxel spacing (isotropic) to standardize the spatial resolution. Noise reduction as in processing such as Gaussian smoothing or median filtering to reduce noise in the images. Normalization is to intensities to correct for variations in image acquisition parameters and scanners. Common methods include Z-score normalization or min-max scaling [21][32].

Step4: Feature extraction: It defined the process of transforming complex data into simplified and meaningful characteristics for analysis and modeling. It involves selecting and generating the relevant features from the initial data to enhance model execution and interpretability. Feature extraction

facilitates reduced dimensionality, eliminates noise, and highlights valuable information, making it simpler for algorithms to learn and make predictions [32].

Step5: Feature selection: Specifically, the feature selection is to choose a subset of the most relevant features from a larger set of available features in a dataset. The goal is to reduce the number of input variables while preserving as much valuable information as possible. Feature selection is crucial in machine learning to improve model efficiency, reduce overfitting, and enhance model interpretability. It also speeds up training and prediction processes. Various techniques, such as filter and embedded methods, can be employed for feature selection [32].

First import the feature extractor model which is the only model needed to extract the radionics toolbox, second store the location of the image and the segmentation where we want to extract features from. Additionally (we are not needed but used with rnnd files, we specify a location for the param file We customize the extractor which can be implemented in different ways, the first way is to use the default settings contained in the toolbox. The second is to define the parameter with specific classes within the dictionary for example set the (binwidth =20), (Sigma=1,2,3), and (verbose= true). Then initiate the extractor. The third method is done by passing the location of the parameters which will be loaded and used by various parameters. Finally, extract the features by calling the extractor and then showing the results. The result contains general information containing information extraction to enhance reproducibility, the rest of the results are precalculated features, and every feature name is defined by the filter, feature class, and the feature name followed by its value [32].

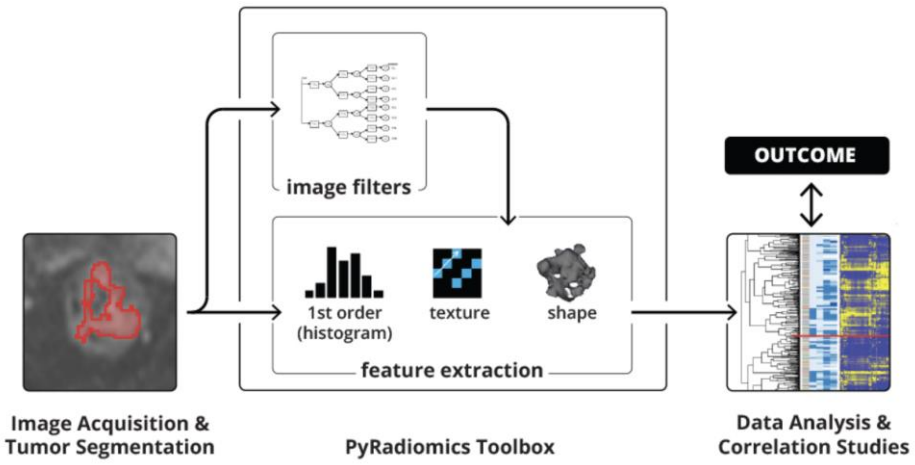


Figure 3.2 Structure of Pyradiomics [33]

The concept of explainability in machine learning (ML), specifically in detecting shape features in MRI images, is fundamental for several aspects, explainability ensures that the ML algorithm's decisions are apparent and understandable. In the circumstances of medical diagnosis from MRI images, healthcare professionals and patients must have accurate and trusted outcomes. The detected feature in

our study the shape features in 2 and 3 dimensions are fundamental for answerability. In the medical field, explainability specific provides medical experts the ability to understand and validate the results. ML can identify the errors or biases in the feature detection process. If an algorithm consistently misinterprets certain shapes, medical practitioners can intervene to correct or fine-tune the model. Also, it gives insights into the value of different features and whether can they aid the final decision and help prioritize certain image acquisition techniques for reliable diagnostic results. Using ML models in a clinical setting, medical professionals may need to interact with the model's decisions. The explainable system allows one to identify and hypothetically dominate or adjust the model's proposals. To achieve explainability in ML for detecting shape features in MRI images, we can employ various techniques:

1. **Feature Visualization:** Visualizing the detected features in the MRI images for interpretability. This can include highlighting regions of interest (ROI) or overlaying distinguished shapes on the initial images.
2. **Feature Importance:** Presenting the individual features in the model's decision. Methods like feature significance scores or saliency maps can help in this interest.
3. **Model Interpretation Tools:** Utilize dedicated model interpretation tools such as Pyradiomics to facilitate in-depth analysis and understanding of the model's workings.
4. **Clinical Validation:** Collaborate closely with medical experts to validate and interpret the identified features within a clinical framework.
5. **Documentation:** Provide a lucid and comprehensive explanation of the training procedure, model structure, and assessment criteria.

By improving the explainability in ML for detecting shape features in MRI images we can improve the trustworthiness and usability of the system in clinical practice while ensuring that it matches with medical expertise and regular requirements [21][32].

3.6 Machine learning algorithms

Machine learning is the field of study within artificial intelligence that empowers computers to learn and improve their performance without explicit programming. Arthur Samuel's Definition (1959):

> *"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."*

The key element factor for developing effective ML models [34][44], is to understand the data that will be trained as an input to the model:

1. Supervised learning: In supervised learning, the algorithm is provided with labels for the training data, where each example has input data and target labels. With this type, the algorithm learns to make predictions or classifications based on this labelled data.
2. Unsupervised Learning: It includes a learning algorithm based on unlabelled data, where the algorithm tries to detect the patterns, relationships, or clusters within the data without explicit guidance.
3. Semi-Supervised Learning: This type combines elements from supervised and unsupervised learning. It uses small amounts of labelled data with a larger amount of unlabelled data to improve model performance.
4. Reinforcement Learning: This is a type of ML agents learn to make decisions based I the environment. It receives feedback in the form of rewards or penalties based on its actions and according to that feedback, it will improve its decision-making every time [34][44].

In the context of prostate cancer diagnosis using MRI and radiomics features, implementing different ML algorithms is crucial for comprehensive analysis. This approach will allow for the comparison model implementation, ensuring the selection of the most suitable algorithm for the task.

3.7 Supervised Machine Learning Algorithms

1. **Random Forest**: It is an ensemble learning method. This ML Combines multiple decision trees to form a "forest". Each tree is trained on a random subset of the data and features. Reducing overfitting is the main advantage between RF over DT. trained via the bagging (sampling with a replacement) or pasting(sampling without replacement) method The final classification depends on the majority vote from all trees. Handles both continuous and categorical data well, is less prone to overfitting, good for large datasets. Widely used in classification problems like medical diagnoses [35].
2. **Decision Tree Classifier:** Tree-based classifier. However, it is constructed of a tree structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Easy to interpret and visualize, can manage both numerical and categorical data. This classifier is widely used in customer segmentation and business decision-making [35].
3. **Logistic regression:** This classifier is Regression analysis, mostly used to model the probability of a certain class or event. It uses a logistic function to model a binary dependent variable. LR provides probabilities for outcomes, efficient and easy to implement. Credit scoring, medical fields (e.g., likelihood of a disease). In equation 2.3, we have a logistic regression model represented in a vectorized form. The logistic function, also known as the sigmoid function denoted as $\sigma(\cdot)$, is used in this context. This sigmoid function is characterized by its S-shaped curve and produces output values in the range between 0 and 1.

$$p^{\cdot} = h\theta(x) = \sigma\,(\theta T \cdot x) \qquad \text{…….(1)[36].}$$

From equation, σ represents the sigmoid function, which transforms the linear combination of the model's parameter vector θ (including bias terms) and the instance's feature vector x into a probability estimate $p^{\cdot}$. This probability estimate indicates the likelihood or probability of a binary event or outcome associated with the given input features x [36].

4. **SVC classifier:** Support Vector Machine (SVM) for classification. Find the hyperplane that best splits up different classes by extending the margin between them. Effective in high-dimensional spaces, memory efficient. It is implemented in applications like Image classification, and text categorization [35].
5. **Ridge Classifier:** Regularized Linear Regression. Extends linear regression by adding a regularization term. This helps to reduce model complexity and prevent overfitting. Reduces multicollinearity in regression, a more stable solution than simple linear regression. Situations with high multicollinearity, prediction in genetics [35].
6. **Naive Bayes classification:** This classifier is based on Bayes theory and assumes the independence of predictors. It's simple and effective, particularly for large datasets. Despite its simplicity, Naive Bayes classifiers can outperform more complex models and are widely applied in text classification [35].
7. **K-Nearest Neighbour classification:** It is a simple machine learning algorithm for classification. It assigns the class label to a new data point based on the majority class among its K nearest neighbours in the training dataset. K – represents a user-defined parameter. This algorithm is easy to be used but it can be sensitive to K value and the choice of the distance metric [37][38].

3.8 Cross Validation

Cross-validation is a technique for assessing how machine learning models generalize to an independent dataset, used to prevent overfitting. It involves dividing the data set into several subsets, repeatedly training the model on some subsets while validating on others and averaging the results to get a more accurate measure of model performance. This method is particularly useful when dealing with limited data, as it maximizes the use of available data for training and validation. Bellow cross-validation works and it's beneficial:

1. Partitioning: Split data into k subsets (folds). A common choice is k = 5.
2. Model Evaluation: For each unique group, the following process occurs, The model is trained on k-1 folds. Remaining fold is used as a test set to evaluate the model.

3. Iteration: This process is repeated k times; each fold being used once as the test set.
4. Aggregation: The k results from the folds averaged to produce a single estimation.

In the model, we implemented cross-validation with 5 folds and 10 folds. Also, we have extracted the hyperparameter for comparison. The machine learning model's performance was measured by using a five-folds cross validation strategy. That's related to assessing the generalization of ML model to an independent dataset and mitigating the overfitting.

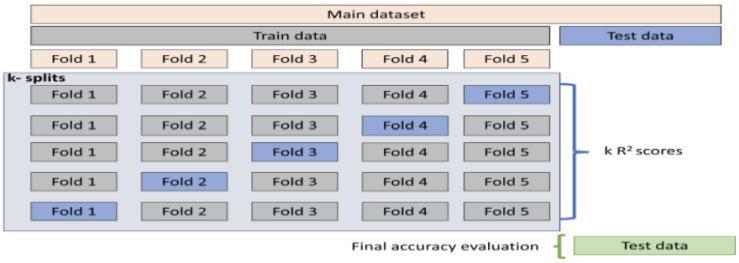


Figure 3.3 Description of 5-fold cross-validation[39].

The final performance of the assessment involved comparison between the model's accuracy on the untouched test set against the training confidence intervals like the work in reference [39].

3.9 Software

In this project, the technical framework is based on Python 3.10.7, the dynamic and high-level programming language used for its readability and efficiency in software development.

3.9.1 **Numerical Python:** NumPy, a fundamental library in Python, is employed for advanced scientific computing and data analysis, particularly with numerical data and multi-dimensional arrays.

3.9.2 **Scikit-learn**: Scikit-learn is a versatile, open-source machine-learning library that facilitates both supervised and unsupervised learning. Known for its comprehensive collection of built-in algorithms and models, scikit-learn enables efficient training with its straightforward 'fit 'method.

3.9.3 **Radiomics:** Alongside Python, the Radiomics library, with Python version 3.7.0, plays a pivotal role in this project. Radiomics is a specialized library in Python tailored for extracting many advanced quantitative features from medical images, particularly in the field of radiomics. This library is crucial for processing and analyzing MRI scans.

Chapter 4

Data and Methods

4.1 Dataset

This Dataset involves data collected from a cohort of 454 participants through T2-weighted measurements. The primary focus of this research is to assess the demographic profiles of individuals who underwent biopsy procedures and underwent MRI image characterization.

T2-weighted (T2w) axial sequences from participants were sourced from the PI-CAI (Prostate Imaging: Cancer AI) initiative spanning the years 2012 to 2021. Along with these sequences, accompanying whole gland masks of the prostate were also procured. The imaging was executed using Siemens or Philips imaging apparatus. The acquired images had an in-plane resolution of 0.5 mm by 0.5 mm and a slice thickness of 3 mm, facilitated by a surface coil. The inclusion of MRI sequences in the research was predicated on biopsies that were confirmed either systematically through MRI guidance, or a fusion of both methodologies. These were categorized as csPC (Gleason score ≥ 7) or ncsPC (Gleason Score < 7 or =0). Throughout this documentation, ncsPC is regarded to encompass both biopsy-verified low-risk PC and negative analysis outcomes. The images were then filtered according to the value of the Gleason score from Demographical clinical features into cancer and healthy. Applying radiomics extraction features for further study statistical characteristics with ML algorithms and comparing the most relevant metrics such as ROC-AUC, confusion matrix, and accuracy.

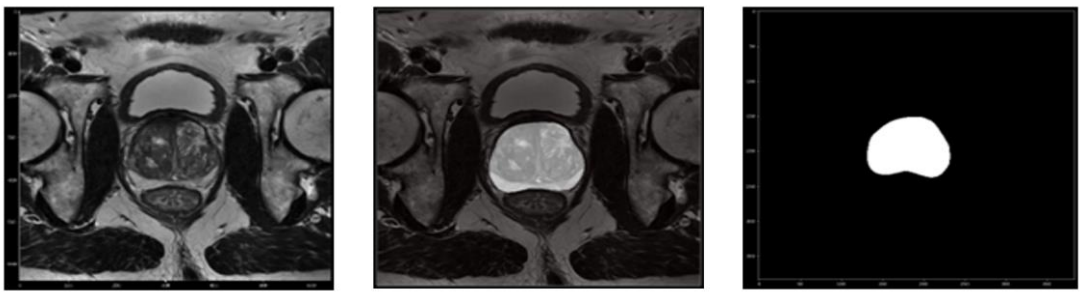


Figure 4.1 Axial T2-Weighted Scans with Cancer Lesion Segmentation for prostate

In In the figure shows 'a' standard MRI, 'b' displays the same region with a highlighted area, likely indicating an abnormality or area of interest, and 'c' is a binary mask.

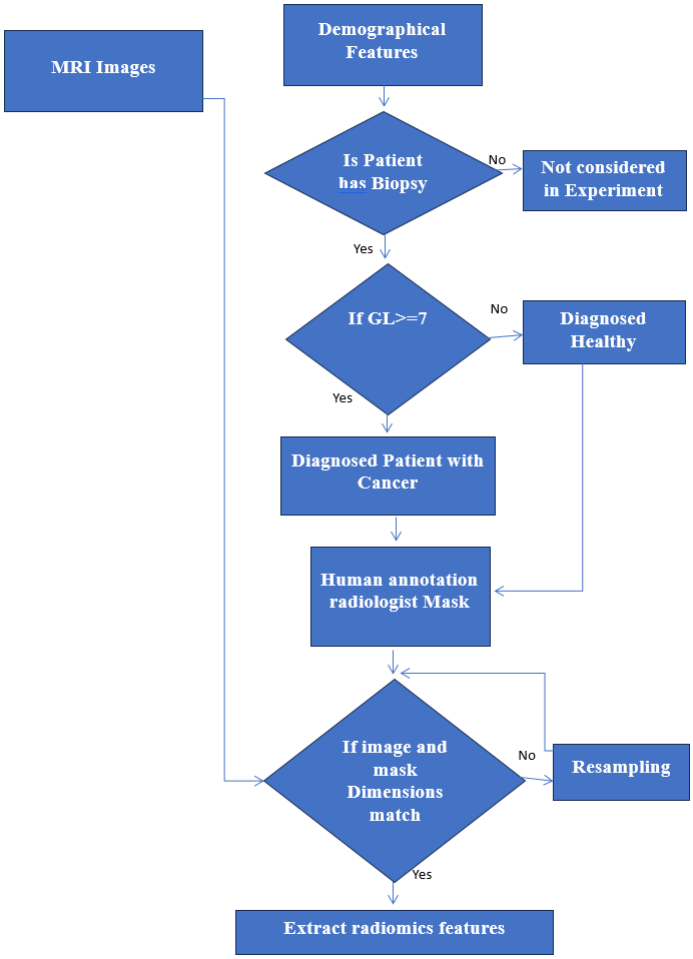


Figure 4.2 Approach to patient filtering

4.2 Data Preprocessing and Model Preparation

Our approach involved adjusting MRI images to a uniform scale with dimensions of 128x128 and normalizing intensity values to a [0, 1] range. Each image was then paired with its corresponding biopsy results to validate the findings. A Gleason score of 7 or above indicates clinically significant cancer (cS), while a score below 7 is considered not significant (ncS), as shown in the table below.

Non-Cancerous Significant (ncS)	217
Cancerous Significant (cS)	237
Total	454

Table 4.1 Represent distribution of patients.

The features extracted for patients to assess the implemented machine learning model's robustness, we divided the dataset into training and testing subsets at 80/20 ratio. This split was carefully executed at the patient level, avoiding data leakage, and preserving the evaluation's integrity.

Consequently, we allocated 363 sequences for training and reserved 91 for testing the algorithm's performance. As illustrate in the figure4.3. and table 4.1. The machine learning models were trained to predict the diagnosis of the patients from MRI images.

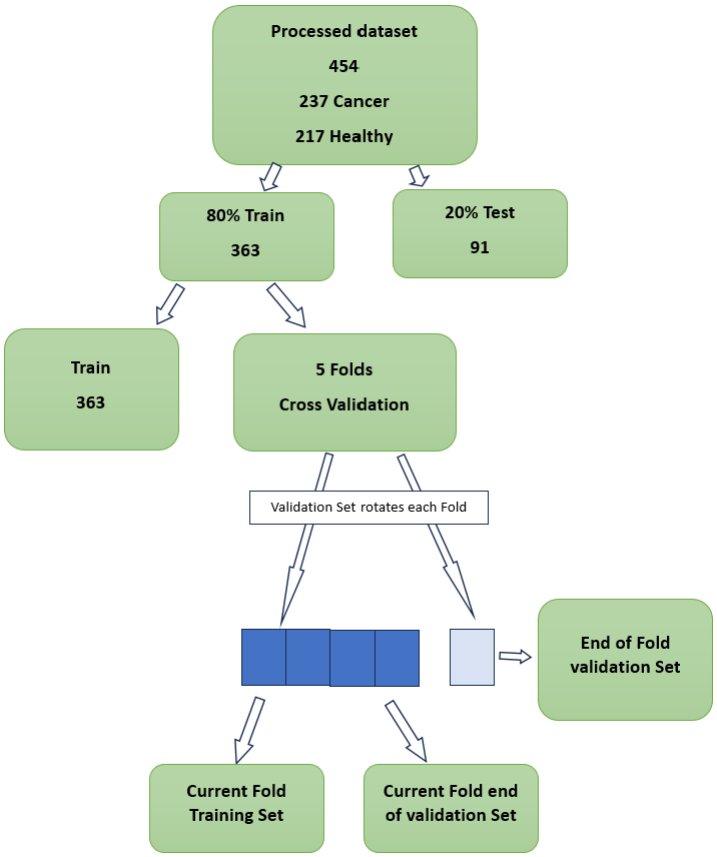


Figure 4.3 Data Partitioning and Validation Flowchart for processed dataset

4.3 Radiomic Feature Extraction and Selection

We assessed the performance of two scenarios implemented with Radiomic feature. In figure 4.4, implemented computer aided diagnosis CAD framework based on MRI images used to detect Radiomics, the workflow commences with image processing, encompassing pre-processing, segmentation, and MRI image registration. Once these steps are completed, the images are utilized in CAD. This involves selecting the region of interest (ROI), followed by feature detection using

pyradiomics. Subsequently, feature extraction takes place, and the extracted features are classified using machine learning (ML) techniques[1][3].

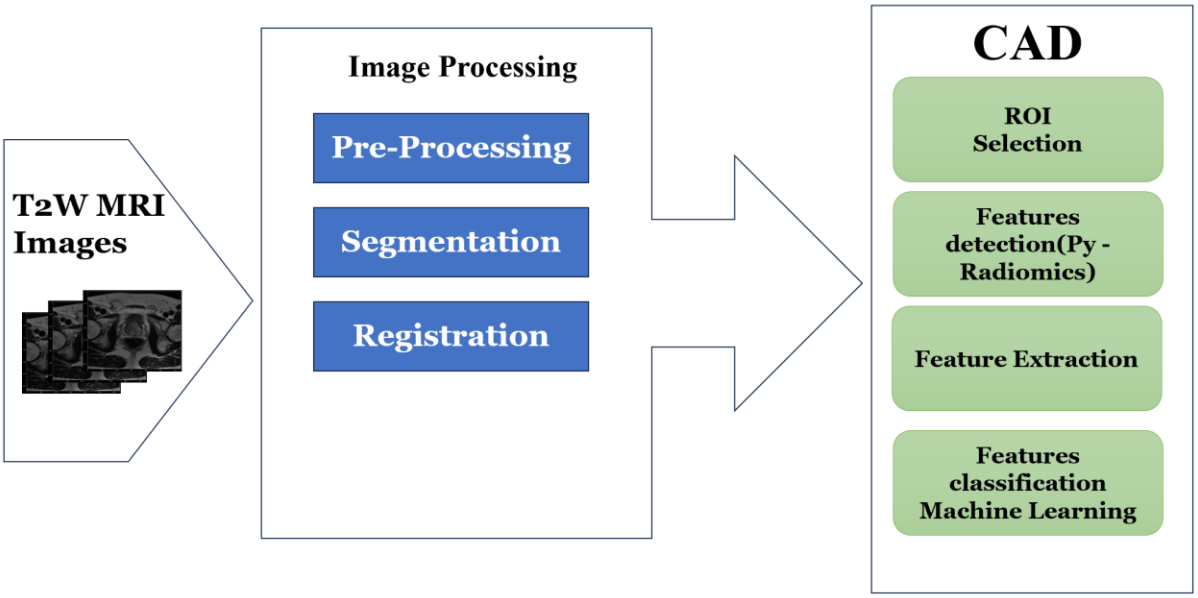


Figure 4.4 Implemented CAD framework

The implemented feature extraction is done by an open-source PyRadiomics python package for the extraction of Radiomics features from medical images. The extracted features are 131, we focused on selecting shape features in 2D and 3D. The elected features for model evaluation from both 2 and 3 dimensions are (13 features) those features elected based on several medical research [1], as the shape features help to distinguish cancerous from normal tissues, as tumors tends to be irregular shapes that can refer to presence of PC. Then, we predict and compare with the test for accruing evaluation metrics [2].

4.4 Machine Learning Modules

In the preceding chapter, Section 3.7 delved into various categories of machine learning techniques. The choice of those algorithms is based on their ability to provide robust, interpretable, and crucial results for medical decision-making. The training process and subsequent evaluation are guided by the following principles:

- The data has undergone a thorough cleaning and preprocessing to be prepared for ML models.
- The dataset divided into training and testing sets to ensure a robust evaluation.
- The training subset is used with the k-fold cross-validation technique to tune the model.

Following the training process and once the model is tuned in, each model undergoes evaluation using the test set. Figure 4.5 illustrates the training and evaluation of the baseline, and sheds light on

the optimization process. Furthermore, after models are established, we processed with creating multiple optimized models by leveraging the hyperparameter package in Python. In subsequent sections, the evaluation process encompasses a range of metrics, including accuracy, precession, recall, F1-score, and receiver operating characteristics- area under curve (ROC-AUC).

4.5 Radiomic Shape Features

We implemented feature selection that chooses the top 13features uses them to train the ML Then, it evaluates the model's performance using various metrics, including accuracy, ROC-AUC, precision-recall curves, GINI, and a confusion matrix. The main reason is for detecting the most relevant features for our classification task and building a predictive model based on those selected shape features.

Cuocolo et al. in [1] employed multivariable logistic analysis to extract shape features, which were then analyzed to determine their association with clinically significant prostate cancer (csPCa). The effectiveness of these diagnostic methods was evaluated by calculating the area under the curve (AUC) in the receiver operating characteristic (ROC) analysis. The same we implemented with shape features summarized the table below with the importance of the selected features [1].

shape Feature Name	**Definition**
Elongation	Elongation shows the relationship between the two largest principal components in the VOI shape.
LeastAxisLength	The smallest access length of the shape, which can assess to differentiate between normal and abnormal prostate shapes, where abnormal may refer cancer.
MajorAxisLength	This feature yields the largest axis length of the VOI-enclosing ellipsoid
Maximum2DDiameterColumn	The largest diameter across the shape in a given 2D slice. Significant changes can reflect growth within the prostate.
Maximum2DDiameterRow	The largest diameter within shape in the row orientation of the 2D slice. Significant changes can reflect growths within the prostate.
Maximum2DDiameterSlice	The largest diameter on any 2D slice of the shape. This can help in detecting tumours that may be not informally shaped.
Maximum3DDiameter	Maximum 3D diameter is defined as largest pairwise Euclidean distance between tumor surface mesh vertices
MeshVolume	The total volume within the surface mesh of the shape. It can indicate enlargement or irregular growth within the prostate.
MinorAxisLength	The length of the smallest axis of the shape. The difference in shape can refer to abnormalities in the prostate.
Sphericity	Measure of the spherical or roundness of the shape of the tumor region relative to a sphere. The tumors cause to prostate lose its roundness shape.
SurfaceArea	The total sum of the area of each triangle in the mesh. An increase can indicate enlarge prostate or surface irregularities because of tumors.
SurfaceVolumeRatio	The ratio of surface area to volume can indicate changes in shape and texture indicative of cancer.
VoxelVolume	The volume was measured in voxels with MRI. Abnormal volumes can refer possibility of cancerous growth.

Table 4.2 Py-Radiomics Shape features [1].

The features in Table 4.2 that were used in the implementation of this study are important in detecting the PC and may reveal changes in the size, shape and texture of the prostate that are characteristics of the presence of tumors [1]. We evaluate extra features like Elongation, Sphericity and the rest listed in Table 4.1 to determine whether those features reflect the presence of PC.

Modern radiomics, which these features are part of, allows more quantitative analysis of imaging data and can lead to more accurate diagnosis of the PC. In our work we extracted (MeshVolume, SurfaceVolumeRatio, VoxelVolume) as it is affected by the presence of tumors in the prostate while Cucolo extracted the same shape features with one volume for the prostate. Radiomic shape features extracted from MRI can significantly enhance the detection of clinically significant prostate cancer.

4.6 Experiment Setup

The data extracted from the MRI T2weight classified according to Gleason score from demographical clinical features, the number of patients 454 with 217 healthy (GS <7) and 237 cancer patients with (GS>=7) the extracted features is 131, in this project, we aim to analyze the shape features. Thus, we extracted the Py-Radiomc features in 2D, and 3D. Two scenarios were used with the extracted dataset, the first implementation with scenario 1, used the unaltered extracted data, which exhibits symptoms of underfitting. I intend to employ this raw dataset as a baseline for comparison against an alternative scenario. In the second scenario, various data resampling techniques were employed, including random under sampling, Near Miss, and SVMSMOTE, to address the underfitting issue and assess their impact on the dataset's performance.

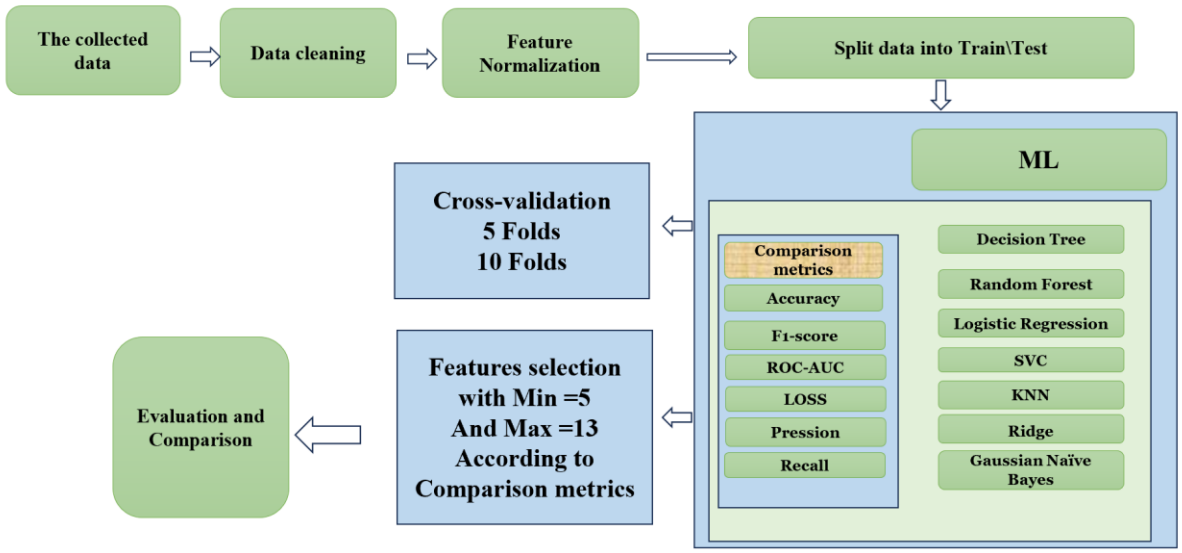


Figure 4.5 Representation of data analysis pipeline

The features extracted are shape features, histogram-based features, and texture features. Besides, texture features covered, grey level cooccurrence matrix (GLCM), run length matrix (GLRLM), and size zone matrix (neighborhood grey-tone difference matrix (NGTDM) features. For more information about of the features, see Appendix C. ML models were executed using default hyperparameters as in figure 4.5, relying on the given feature set without any additional feature engineering or customization.

ML implementation as we can see figure 4.5, with different steps beginning from the data collection from e MRI images using features extraction Py-Radiomics, then cleaning data, this involves correcting inaccuracies, handling missing values, and other data quality issues. In Feature normalization, the features are scaled or distributed, which is important for models that are sensitive to the scale of data. The normalized data is then split into train and test 80%, and 20%. In Cross-validation the training data is further divided into subsets (folds) to validate the models during training. Five folds and ten folds, which are to estimate the performance of the model on unseen data and repeated three times. Five-fold cross-validation balances descriptive power and computational efficiency for our dataset, reducing the risk of overfitting and performance overestimation that we observed with ten-fold cross-validation on our fixed size dataset [40][41].

Comparison Metrics: Accuracy, F1-score, ROC-AUC, Loss, Precision, and Recall.

ML Algorithms: Several algorithms are applied, including Decision Tree, Random Forest, Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbours (KNN), Ridge Regression, and Gaussian Naive Bayes.

Feature Selection: Minimum of 5 and a maximum of 13 features. The selection is based on the comparison metrics mentioned above.

5.6.1 Scenario 1 (Implemented without resampling)

Methods: In this scenario, implemented ML on extracted features without augmentation techniques like resampling, and evaluate models with cross -validation. We aim to assess the performance of the model or analysis based on the raw, unaltered dataset. This can serve as a baseline to understand how well the model performs without any intervention to address class imbalance or other potential dataset issues. This model was implemented with extracted shape features and classified with seven machine learning models, RF, DT, LR, SVC, KNN Gaussian Naive Bayes and Regression and evaluated with cross-validation. The features extracted and evaluated with different metrics. In scenario 1, we avoid under-sampling and prevent discarding some data (missing filled with nan). The reduced dataset may not fully represent the original distribution of patients in the real world. However, this trade-off is acceptable when analysis or visualization requires class balance. This will be tested with the next scenario.

5.6.2 Scenario 2: Enhanced Features with Data Augmentation and Visualization

This scenario is implemented with resampling techniques to handle imbalanced datasets by using SVMSMOTE (Synthetic Minority Over-Sampling Technique) for data augmentation [7], by synthetically creating new samples in the dataset. I also implemented Multi-dimensional Visualization using T-SNE (t-distributed Stochastic Neighbor Embedding) to visualize the high-dimensional data in a lower-dimensional space. The implementation of Multi-Dimensional Visualization using (T-SNE), to reduce the complexity of the data into simpler 2D, and 3D visualization for better understanding the pattern, understanding the relationships between the data and cluster identification with anomaly detection [40].

The dataset obtained from extracting radiomic features of MRI T2W images presents class imbalance, comprising two classes: "Cancer" (1) as the majority class and "Healthy" (0) as the minority class. To address imbalance, we employed resampling techniques, namely "Random Under-sampling" and "Near Miss Under-sampling," along with an resampling technique known as "SVMSMOTE" [7][40][41].Then we select the best techniques according to highest accuracy for models.

- **Random Under-sampling**: Removal of examples from the majority class in the training dataset to achieve a more balanced class distribution.
- **Near Miss Under-sampling**: selects majority class examples based on their proximity to minority class examples, ensuring better balance.
- **SVMSMOTE**: This oversampling method utilizes an SVM classifier to identify support vectors and generate synthetic samples, considering these support vectors.

➢ **Baseline:**

- Exploring the Data (Visualization, Statistics, Distribution for Cancer & Healthy)**.**
- Apply machine learning classifier for the imbalanced dataset before sampling and check the results evaluate through cross validation.
- After applying Sampling, fit the data on ML models and evaluate.
- Apply cross_val _score for three Machine learning classifiers. (LogisticRegression, Decision Tree Classifier, KNeighborsClassifier). We used 5 folds which repeated 3 times and the ROC_AUC as a scoring metric.

➢ **Splitting Data and scaling:** Split the data with an 85:15 ratio that's mean 85% for training set and 15% of the dataset for testing set. Scaling for both Volume and Diameter columns.

➢ **Handling Imbalanced data:**

we use techniques like Random Under sampling and Near miss, SVMSMOTE.

4.7 Comparison and Analysis:

By comparing the results from both scenarios to evaluate the impact of data augmentation and advanced visualization techniques. We also assess the visual differences in class separation and clustering through t-SNE visualization [40].

Five-fold cross-validation provides a good trade-off between bias and variance when estimating model performance. Using fewer folds may lead to high variance in performance estimates, more folds may introduce bias due to smaller training sets. Ten-fold cross-validation was also implemented, however, the data set is small, this may yield seemingly elevated accuracy metrics due to the limited amount of data allocated to each fold. Therefore, we assess it with 5 folds. In our analysis risk factors are considered for the importance of model selection. In table 5.1 Type II errors, recognizing the critical impact of failing to identify true risks. Following, Type I errors hold a level of importance, as incorrectly identifying a non-risk can still have significant consequences [5]. The accuracy of the model ensures that the overall proportion of correct predictions is achieved. Finally, the ROC-AUC score is considered, which reflects the model's ability to discriminate between the classes across various thresholds. Also, we evaluated the confusion matrix as we examined the false positive and false negative. Multiple metrics allow for a wide view evaluating models [41][42].

Chapter 5

Results and Evaluation

In this chapter, we focused on a critical analysis and evaluation of the results from two implemented scenarios applied to datasets extracted from MRI images.

		Positive (1)	Negative (0)
Predicated value Positive/Negative	Positive (1)	True positive (TP)	Type I Error (FP)
	Negative (0)	Type II Error (FN)	True Negative (TN)

Table 5.1 Confusion matrix [44]

5.1 Scenario 1 evaluation:

In scenario 1 7ML are implemented, for analysing the model performance we extracted several parameters such as accuracies, loss, precession recall, F1 score, and ROC-AUC. From Figure 5.2 the confusion matrixes reveal Naive Bayes minimizes Type II errors, while Logistic Regression, Ridge, and KNN exhibit higher values.

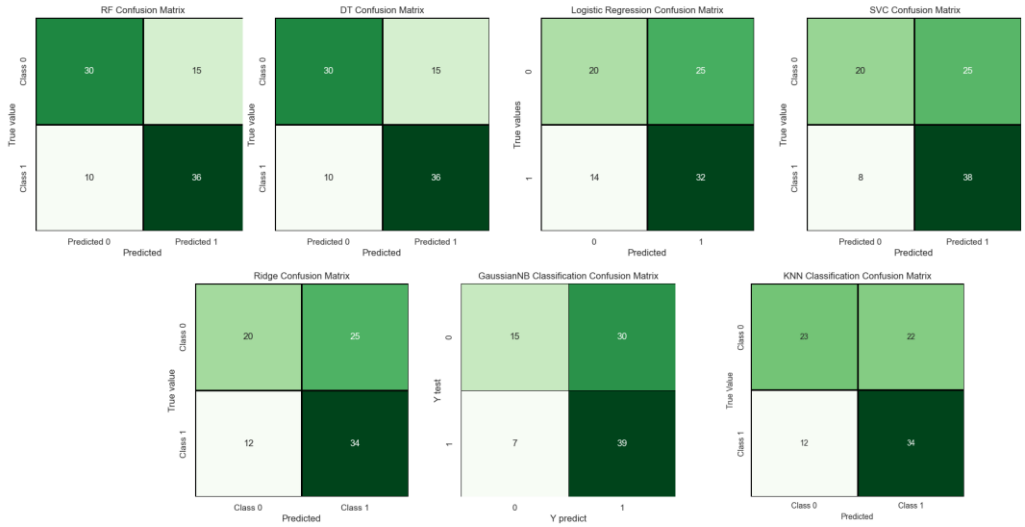


Figure 5.1 Scenario1 confusion matrix comparison

The value of mean for both train and test sets for the extracted shape features, showing consistency across most features, indicating a well-balanced dataset. However, The standard deviation values for Sphericity (0.023, 0.024) in train and test and test respectively are low, higher SD in 'Mesh volume' and 'Voxel volume' especially in testing sets. For detailed results available in Appendix B.

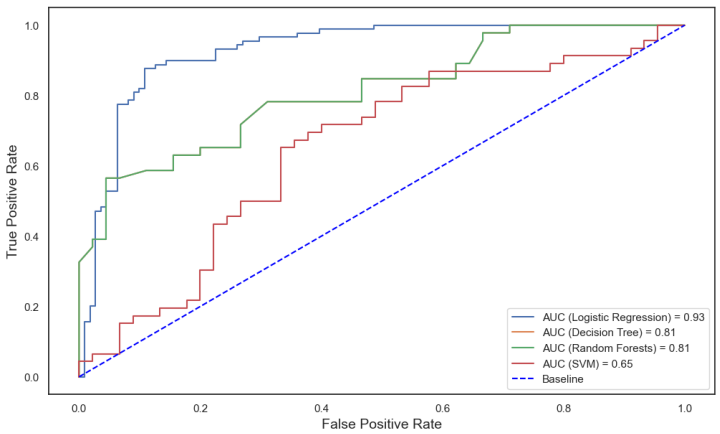


Figure 5.2 The ROC-AUC values for the highest value of implemented ML in Scenario1

The ROC curve Figure 5.2 highlights LR superior ability to distinguish between classes with highest ROC-AUC (0.93). LR shows consistent high performance across several metrics, with notable stability in model predications reflected by standard deviation above 0.06.

Metric	LR	SVC	Ridge	DT	RF	K-NNN	Naive Bayes
Accuracy:	0.88	0.63	0.59	0.73	0.73	0.62	0.59
Precision	0.71	0.60	0.57	0.71	0.71	0.60	0.56
Recall	0.78	0.82	0.73	0.78	0.78	0.73	0.84
F1 Score	0.86	0.697	0.64	0.74	0.74	0.66	0.67
Log Loss	0.37	0.66	0.66	0.52	0.52	13.64	1.043
Gini Coefficient	0.86	0.30	0.24	0.17	0.62	0.25	1.18
ROC-AUC area	0.93	0.65	0.62	0.81	0.81	0.68	0.61

Table 5.2 ML implementation Scenario1

We can notice from the Table 5.2, ML implementation results the LR excel in accuracy (0.88) and F1 score (0.86), referring to effective case classification and strong balance between precession and recall. RF and DT lead in precession (0.71), minimizing the FP. Naïve Bayes achieves the highest recall (0.84), suggesting it effectively minimize the FN but may overclassify positives. LR also shows the lowest loss log (0.37), reflecting reliable probability predications. However, Naïve Bayes high Gini Coefficient signals potential overfitting concerns. The ROC-AUC scores corroborate LR discriminative ability making it the most robust model among those evaluated.

5.2 Cross validation:

The cross-validation results show the model with consistent and high performance, indicating by mean scores above 0.88 in accuracy, precession, recall, and F1 score, and above 0.94 in ROC-AUC. Results as shown in the Table 5.3. The ROC-AUC suggesting excellent ability to distinguish between

classes. Log loss at 0.31, reflecting good confidence in classification tasks. The ROC-AUC suggesting excellent ability to distinguish between classes. Log loss at 0.31, reflecting good confidence in classification tasks. Fold 2 showed the best values regarding all metrics with lowest value for loss.

	Accuracy	Log Loss	Precision	Recall	ROC AUC	F1 Score	Gini
Fold1	0.87	0.32	0.87	0.83	0.92	0.87	0.85
Fold2	0.92	0.23	0.92	0.86	0.96	0.919	0.93
Fold3	0.90	0.32	0.90	0.96	0.93	0.90	0.86
Fold4	0.875	0.30	0.85	0.86	0.94	0.87	0.89
Fold5	0.88	0.26	0.92	0.95	0.96	0.87	0.92
Mean	0.88	0.31	0.89	0.89	0.94	0.88	0.89

Table 5.3 Cross validation scenario1

5.3 Evaluating Scenario 2:

The dataset presents a variety of shape-related attributes, including elongation, axis length, and diameters, each with specific means and standard deviation that suggest a wide range of shape characteristics. The dataset has imbalanced in healthy 217 and the cancer 237. In this scenario we handle imbalance by using Random under sampling, Near Miss and SMOTE [7], then implementing ML. The sphericity attribute average range is around 0.76, hinting at general trends towards spherical shapes within data.

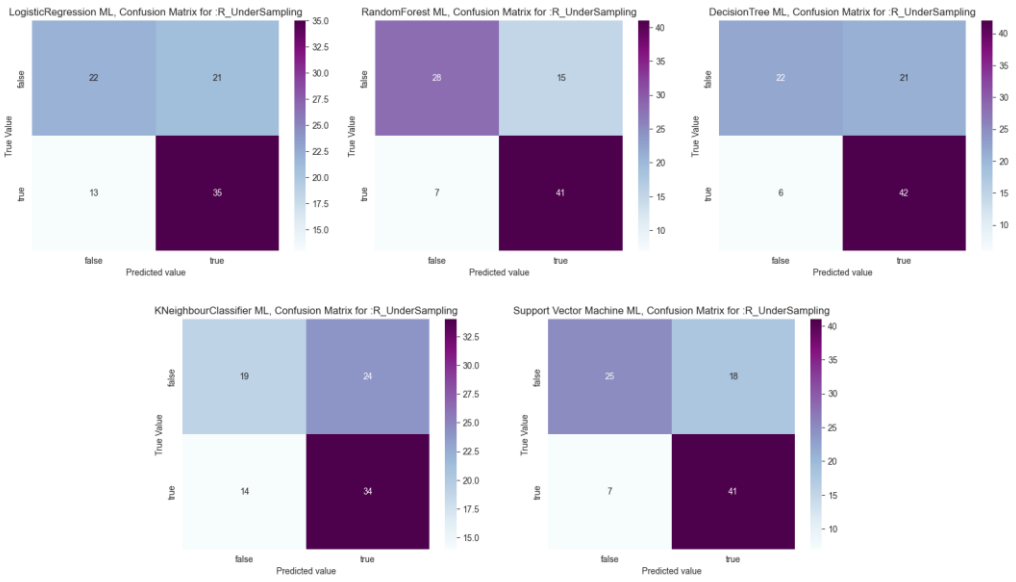


Figure 5.4 Scenario2 confusion matrix

In reviewing Confusion matrix figure 5.4. Random Forest stands for the low value for type I error (FP), and only (15) misclassification as compared to SVC with highest value (24), and Type II error (7) for Random Forest. When considering both Type I and Type II error the RF is best results indicating effective classification with strong ability to distinguish between classes while minimizing miss classification.

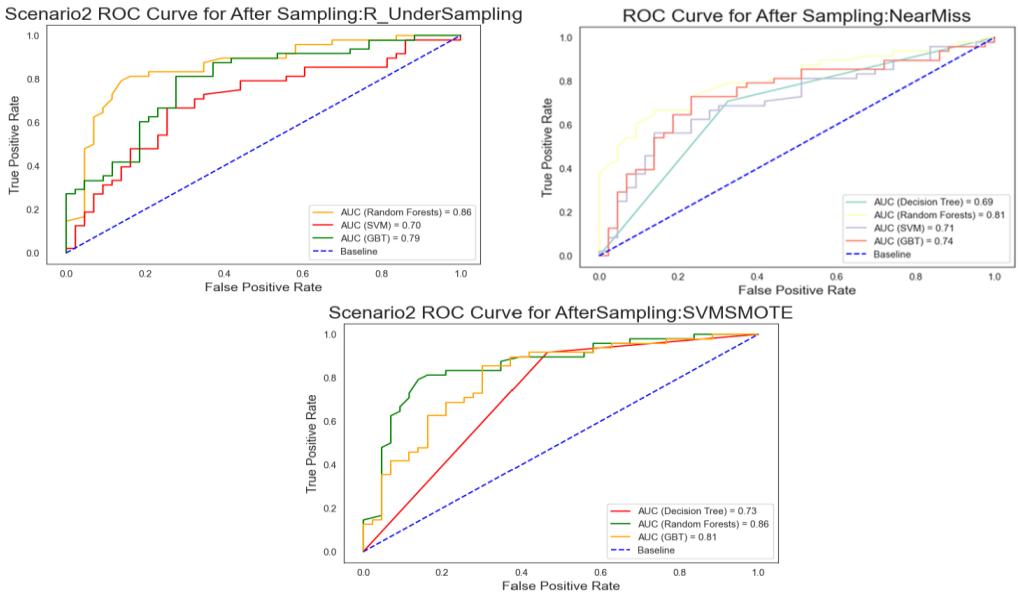


Figure 5.5 Evaluation ROC-AUC for the ML models

For the ROC-AUC graph representation give comprehensive overview of how classifier performance of the scenario2 across all possible thresholds. The ROC in figure 5.5 indicates that the Random Forest (RF) algorithm exhibits strong performance with a high AUC across different sampling methods, underscoring its robustness. Logistic Regression (LR) and (SVM) demonstrate moderate efficiency, with their curves and AUC values suggesting average discriminative capabilities. For more detailed graphs are available for all ML in Appendix A.

	Model	Accuracy_Score	Precision_Score	Recall_Score	F1_Score
1	RF	0.747253	0.711864	0.875	0.74171
3	SVM	0.736264	0.7	0.875	0.729532
2	DT	0.736264	0.693548	0.895833	0.727406
5	GBT	0.714286	0.677419	0.875	0.70469
0	LR	0.637363	0.653061	0.666667	0.637099
4	KNN	0.604396	0.607143	0.708333	0.599324

Table 5.4 ML implementation Scenario2

The table 5.4 RF classifier leads with highest acc 0.74, it has strong balance between pression and recall indicating lower rate in both Type I and Type II errors. LR and K-NN fall behand with lower acc and higher Type I and II errors indicating low recall and pression. The distribution Transformer automates feature transformation to align closer with a normal distribution, optimizing each feature individually. The results for Macro are nearly the same as micro with little difference, so we just depend on evaluating the models with macro calculations.

5.4 Best Feature extraction

In evaluating the performance of the algorithms, we note that the features selected as the best predictors, based on their contribution to the highest ROC-AUC and accuracy scores, are as follows **(Maximum 2D Diameter Row, Mesh Volume, Minor Axis Length, Surface Volume Ratio, Voxel Volume).** In assessment, RF and SVC models has the good performance in balancing the accuracy and the errors rates, with RF has slight edge in discriminative power as indicated by ROC AUC. Naive Bayes stand out minimizing Type I errors which is crucial when false alarm is a significant concern in our case of study. The lower AUC for LR, Ridge and KNN suggests those models are less capable in terms of overall classification performance for this dataset.

Chapter 6

Conclusion and Future Work

6.1 Discussion

Based on the results from the previous chapter in scenario1 provided by the confusion matrix, ROC curves, and tabulated performance metrics, LR shows the highest values in accuracy, precession, recall, and F1 score, referring to its best performance overall in this dataset. RF and DT classifiers show similar performance proposing better class separation ability. However, their accuracy is low with precession as compared to LR indicating some limitations. The SVC exhibits good performance in the recall, meaning it is good at identifying positive cases but has low precision and accuracy.

We have noticed in Feature 'Mesh Volume' the Standard deviation values for train and test, have differences as compared to other shape features. Higher SD in 'Mesh volume' and 'Voxel volume' especially in testing sets. However, The SD values for Sphericity in train and test and test are low. Variability in features like 'Mesh Volume' suggests significant differences in prostate size among individuals. The value of mean for both train and test sets for the extracted shape features, showing consistency across most features.

Overall, the mean values indicate well balanced dataset and crucial for reliability, the varying SD highlight the importance of considering feature variability in model training and validation. The observed difference may require model adjustment or normalization. For detailed results available in Appendix B.

Our analysis inspired by Shanker work[2], focusing on key performance metrics such as accuracy, F1-score, and ROC-AUC. Notably, our Logistic regression model, without resampling, achieved acc of 88%. This to be high accuracy rate of over 90% reported by Stainbus et al [6], in their lymphoma classification study using deep learning. These results from scenario1 are encouraging, suggesting that our model is a current high performing model and is well aligned with our objective in developing an explainable model.

For scenario2, Logistic Regression achieved an accuracy of 63.7%, which slightly lower than 74.7% accuracy by Random Forest model, it is notably close to the 59.7% reported by Shanker [7], who initiate nuclear features to be more predictive than texture features. Despite the superior accuracy of RF, the model LR relatively accepted accuracy coupled with simplicity and interpretability, LR provides clarity of how predication is made, thus encouraging the case for employing LR in clinical settings

where interpretability is key. Krajnc, D. et al. implemented in their work [7] with SMOTE technique, applied RF, SVM, XGBoost and NN model, and achieved the best results with RF and NN methods the acc across the cohort are nearly (0.66%) the accuracy for RF. To uncover the complexities and innate groupings within this high-dimensional space, t-SNE merges as a valuable technique. It excels in condensing the data into 2D or 3D visual representations, maintaining the integrity of the original relationships between data points and revealing clusters that are less discernible in full-dimensional space.

The implementation of the random under-sampling, Near Miss Sampling, and SVMSMOTE sampling techniques suggests differential impacts on model performance. Random under sampling shows consist of metrics across the models. Near Miss Sampling generally lowers precision and F1 scores, hinting at a propensity for models incorrectly predicate the positive classes. SVMSMOTE Sampling seems to improve the recall without compromising other metrics, suggesting it may be superior for balancing class representation. Therefore, SVMSMOTE Sampling stands out as potentially the most active technique for maintaining performance equilibrium across accuracy, precision, recall and F1 score in the tested algorithms, for more detailed results the figures in Appendix **A**. In the figure 'Comparison Metrics for Resampling 'results clearly comparing the implemented ML algorithms on a dataset with Resampling Techniques implemented. Gradient Boosting Tree (GBT) appears as the top, excelling in all metrics, indicating its effectiveness in handling imbalanced datasets. Decision Tree (DT) and Random Forest (RF)show close results, with strong performance but slightly lower than GBT, this is due to overfitting or due to data complexity. SVM, LR, and KNN lag, this suggests that they might not be well-suited for this dataset characteristics or for problems with non-linear decision boundaries and imbalance.

The heatmap comparison for the first scenario displays correlation matrix with most positive correlation among featured, dark is in diagonal confirms perfect self-correlation. The scenaro2 heatmap shows similar trends with same trends with positive correlations, however the strength of these correlation might be less as compared to the first scenario, with alittle difference pre-and post under sampling highlights the impact of this techniques on features relationships. Heatmaps available in Appendix B.

6.2Improvements:

For lower-performing classifiers like SVM and KNN, hyperparameter optimization and feature selection could improve results. For high-performing models like Random Forest, ensuring the model isn't overfitting with techniques such as cross-validation is important.

- Ensemble methods: This method like Easy Ensemble or Balance Bagging Classifier, it combines many classifiers to handle imbalance effectively.
- Cost-Sensitive Learning: modify the algorithms to consider class imbalance by assigning different misclassification costs to different classes.
- Anomaly detection: Treat the minority as an anomaly detection problem that can be more robust to class imbalance.
- Collect more data: gather more data for minority classes in our project (Healthy) to balance the dataset, but it is not possible in our case to ask a patient who is diagnosed as a healthy person to undergo a biopsy.
- Adjust decision threshold: Adjust the decision threshold of our model to increase sensitivity or specificity, depending on the specific problem and its consequences.

The dataset we generated from extraction features can lead to potential issues including module bias towards the majority class (cancer patients) and poor performance on the minority class (healthy)t. The solution in this case is either to collect more data (biopsy)from healthy patients which is not reasonable, or we must consider the following strategies:

- Resampling: Oversampling the minority classes by generating synthetic samples for the minority in our case is the healthy diagnosis, or by replicating existing samples to better balance the class distribution.
- Synthetic class distribution: use techniques such as Minority oversampling Techniques SMOTE, as we implemented in scenario 2 to create synthetic samples for the minority class.
- Different algorithms: choose ML algorithms that are less sensitive to class imbalance.

6.3Best features extraction

Five features are considered as best shape features according to comparison metrics accuracy and ROC-AUC (Maximum2DDiameterRow, Mesh Volume, Minor Axis Length, Surface Volume Ratio, Voxel Volume). In (my opinion) a small number of features is better to influence the diagnosis. These features span various aspects of prostate geometry, suggesting that both the external and internal shape are critical in predicting the outcomes of interest. The prominence of those features in the models with the highest ROC-AUC and accuracy. The use of those features likely enhances the model, leading to better performance metrics and more confident decision-making in a clinical setting achieving our objectives. Results for feature extraction in detail in appendix B.

6.4 Conclusion

The results obtained from classification models indicate satisfactory in performance to detect patients, while some models are relatively poorer performance. Several factors may contribute to this outcome. The dataset size does not provide enough information for model to learn complex patterns, or some shape features might not have linear relationship with outcome.

In the first scenario, Logistic Regression (LR) yielded the most favorable outcomes, a result that can be attributed to the algorithm's inherent simplicity and interpretability. However, in the second scenario, which involved resampling, LR demonstrated reduced accuracy and ROC-AUC scores. This decline in performance can be ascribed to the limited size of the dataset. Given the minor discrepancy in the number of healthy versus cancer instances, the issue of underfitting was not significant, and thus the data did not necessitate sampling. Consequently, the results from the first scenario are deemed more rational and in alignment with the objectives of the study, suggesting that for small datasets with minimal class imbalance, the application of sampling techniques may not be beneficial and could lead to a degradation in model performance.

For datasets of substantial size, the second scenario, which incorporates sampling techniques, is often the preferred approach. Sampling methods can significantly enhance model performance by addressing class imbalance, a common issue in large-scale datasets. They ensure that the learning algorithm receives a balanced representation of each class, thereby improving the generalizability and robustness of the model.

Conversely, in the context of smaller datasets, such as in the present study, the application of sampling techniques may not yield the same benefits. Here, the first scenario, which operates without the intervention of sampling, is advantageous. The modest difference in the class distribution between healthy and cancer cases does not warrant the complexities introduced by sampling. In such instances, the detrimental effects of over-sampling or under-sampling can outweigh the benefits, as these techniques can introduce noise or lead to the loss of critical information, respectively. Hence, for datasets with minimal class imbalances, preserving the original data distribution—as in scenario one—tends to be more effective and produces outcomes that more accurately reflect the underlying data structure. This approach aligns closely with the objectives of ensuring model interpretability and maintaining data integrity.

6.5 Future work

Our goal is to advance the field of medical imaging through the development and evaluation of machine-learning techniques. In this study, we explored 2D and 3D reconstruction for prostate cancer analysis, aiming to discover distinctive characteristics of prostate cancer and identify the most relevant predictors. While we have successfully extracted the Mesh (available in GitHub link [43]) we can see prostate Mesh in the figure below. Also, I extracted. sample of mesh of the prostate and the tumor mesh, the future work will focus on the extraction of mesh features, enhancing our computer-based characterization, particularly in the realm of shape perspectives. Also, we can analyze different Radiomic features like Histogram features and reimplement the scenarios with LR as further work.

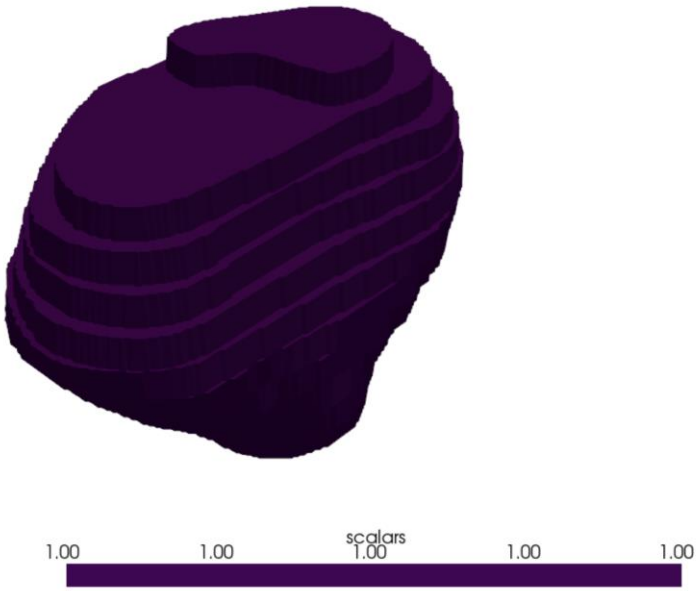


Figure 6.1 Mesh generated for the prostate.

List of abbreviations:

AI: Artificial Intelligence

CM: Confusion Matrix

CT: Computed Tomography

DRE: Digital Rectal Examination

GNB: Gaussian Naïve Bayes

GS: Gleason Score

LR: Logistic Regression

ML: Machine Learning

MRI: Magnetic Resonance Imaging

ncsPC: Non-Clinically Significant Prostate Cancer

PC: Prostate Cancer

PSA: Prostate Specific Antigen

RF: Random Forest

ROC: Receiver Operating Characteristics

ROI: Region of Interest

SVM: Support Vector Machine

SVC: Support Vector Classifier

T: Tesla

T2w: T2-weighted

csPC: Clinically Significant Prostate Cancer

DT: Decision Tree

RF: Random Forest

KNN: K-Nearest Neighbor

FP: False Positive

FN: False Negative

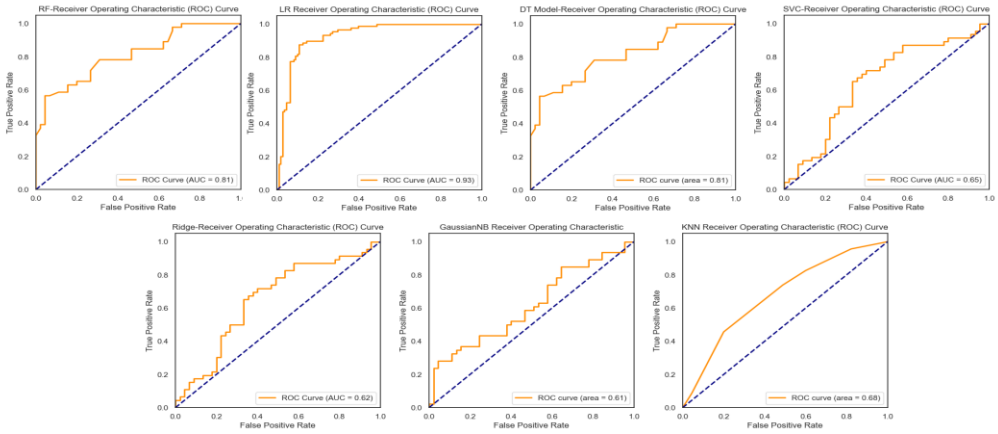
tpr : True Positive Rate

fpr: False Positive Rate

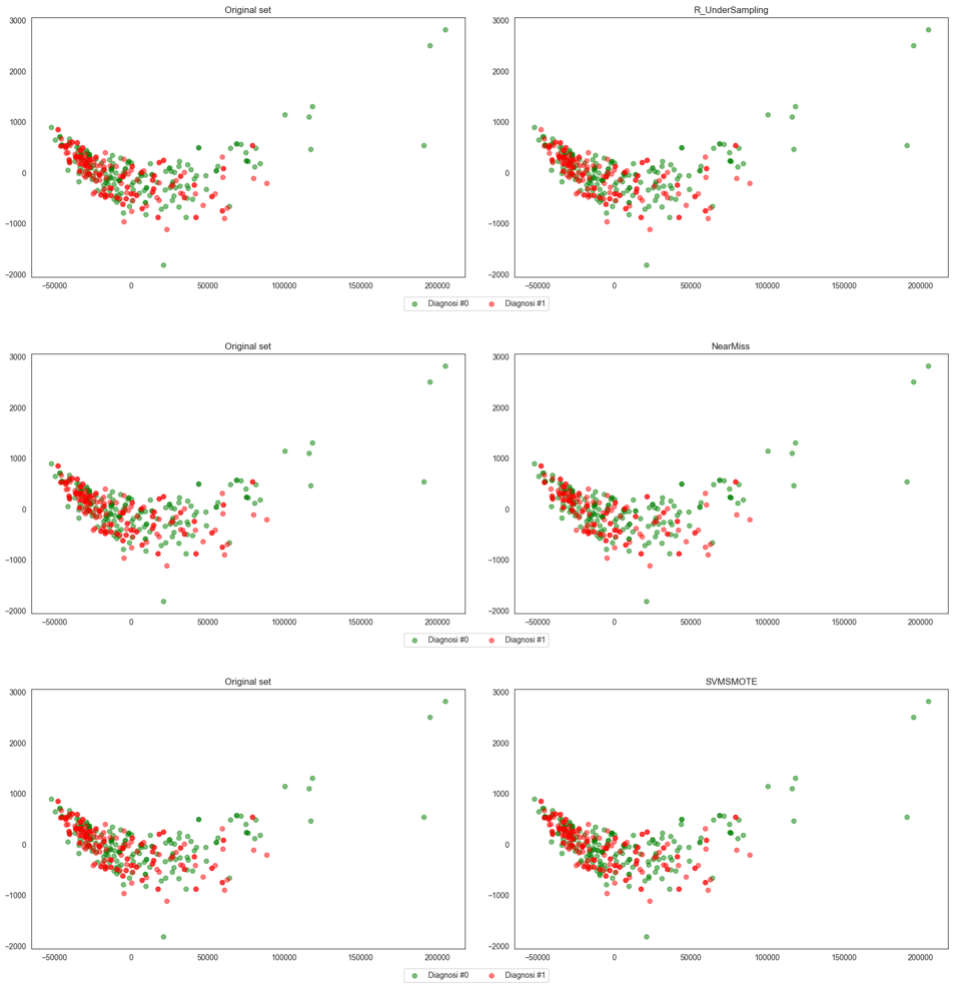
CAD: Computer Aided Diagnosis

Appendix A:

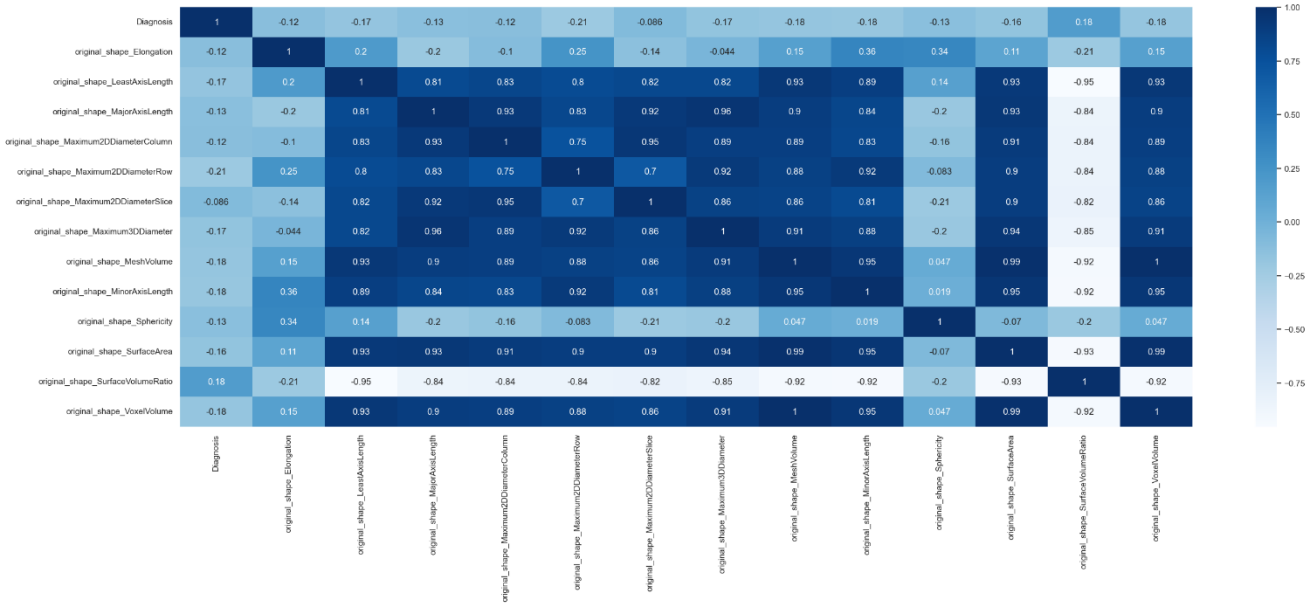
List of Figures



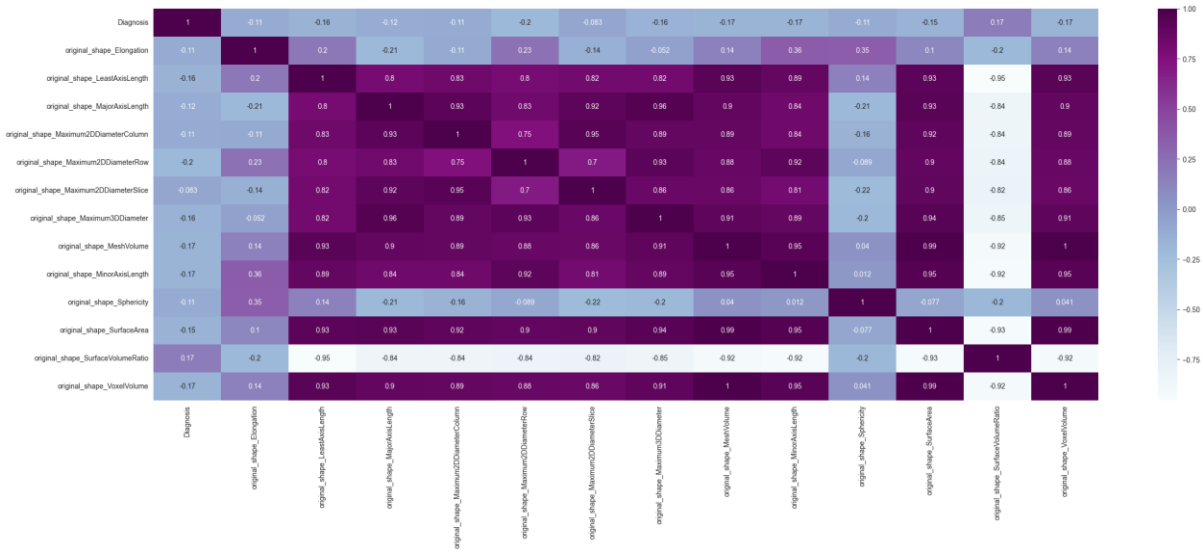
ROC-AUC Scenario1



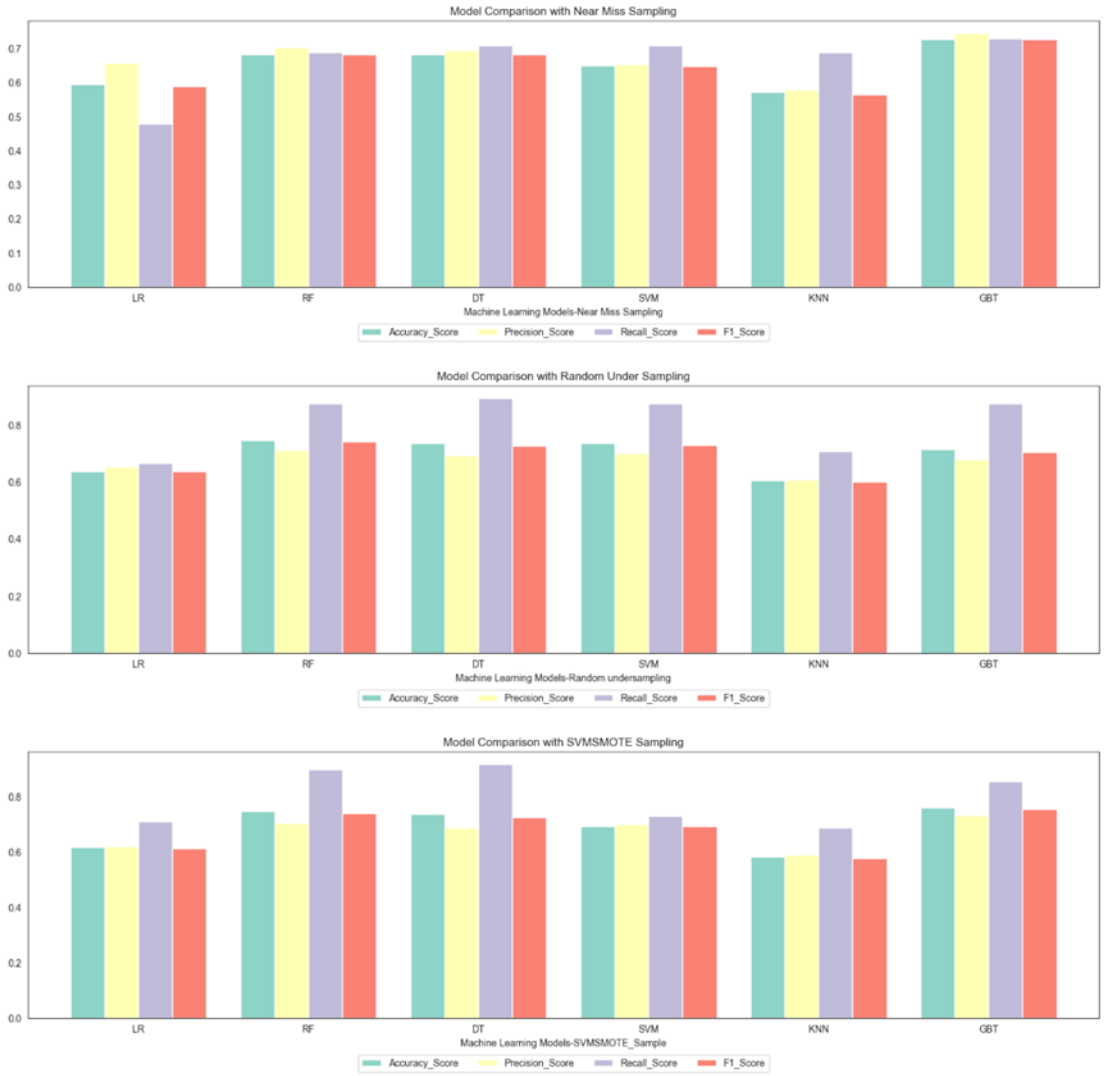
Resampling techniques secnario2



Scenario1 Heatmap



Scenario 2 Heatmap



Comparison Metrics for Resampling

Appendix B

List of Tables

Index	Training Set	Standard Deviation	Mean
1	original_shape_Elongation	0.076468	0.868593
2	original_shape_LeastAxisLength	5.751311	35.47726
3	original_shape_MajorAxisLength	7.24494	49.03069
4	original_shape_Maximum2DDiameterColumn	6.955341	54.82579
5	original_shape_Maximum2DDiameterRow	9.896288	54.43178
6	original_shape_Maximum2DDiameterSlice	7.349865	54.34074
7	original_shape_Maximum3DDiameter	8.822754	58.06406
8	original_shape_MeshVolume	25631.29	54991.46
9	original_shape_MinorAxisLength	6.634372	42.47243
10	original_shape_Sphericity	0.023259	0.763358
11	original_shape_SurfaceArea	2705.671	8969.099
12	original_shape_SurfaceVolumeRatio	0.025583	0.174057
13	original_shape_VoxelVolume	25638.83	55041.53

Table 1. Mean and Standard deviation of the Scenario1 for Train

Index	Testing Set	Standard Deviation	Mean
1	original_shape_Elongation	0.064593	0.879708
2	original_shape_LeastAxisLength	6.287612	35.29461
3	original_shape_MajorAxisLength	7.6387	48.10328
4	original_shape_Maximum2DDiameterColumn	7.744284	54.07042
5	original_shape_Maximum2DDiameterRow	10.15792	53.77554
6	original_shape_Maximum2DDiameterSlice	7.594802	53.43384
7	original_shape_Maximum3DDiameter	9.493782	56.72693
8	original_shape_MeshVolume	29598.11	54408.17
9	original_shape_MinorAxisLength	6.98272	42.24738
10	original_shape_Sphericity	0.024431	0.769559
11	original_shape_SurfaceArea	2989.333	8791.075
12	original_shape_SurfaceVolumeRatio	0.026163	0.174527
13	original_shape_VoxelVolume	29608.17	54459.34

Table 2. Mean and Standard deviation of the Scenario1 for Test

	Accuracy	Log Loss	Precision	Recall	ROC AUC	F1 Score
Fold1	0.87	0.32	0.87	0.83	0.92	0.87
Fold2	0.92	0.23	0.92	0.86	0.96	0.919
Fold3	0.90	0.32	0.90	0.96	0.93	0.90
Fold4	0.875	0.30	0.85	0.86	0.94	0.87
Fold5	0.88	0.26	0.92	0.95	0.96	0.87
Mean	0.88	0.31	0.89	0.89	0.94	0.88
	Accuracy	Log Loss	Precision	Recall	ROC AUC	F1 Score

Table 3 Cross validation 5 folds scenario1

	Accuracy	F1	ROC_AUC	Precision	Recall	LOG_loss	GINI
1	0.86	0.85	0.94	0.89	0.82	0.31	0.88
2	0.89	0.89	0.95	0.90	0.88	0.29	0.90
3	0.91	0.91	0.97	0.90	0.92	0.22	0.95
4	0.93	0.93	0.96	0.96	0.90	0.23	0.93
5	0.92	0.92	0.96	0.92	0.92	0.28	0.91
6	0.85	0.85	0.92	0.87	0.82	0.35	0.84
7	0.86	0.87	0.95	0.83	0.90	0.30	0.91
8	0.91	0.91	0.96	0.90	0.92	0.26	0.92
9	0.94	0.94	0.98	0.96	0.92	0.22	0.96
10	0.85	0.84	0.94	0.91	0.78	0.33	0.87

Table 4 Cross-validation 10folds scenario1

Number of features (RF)	Accuracy	Log Loss	F1 Score	ROC-AUC	Precision	Recall
5	0.703297	0.556757	0.737864	0.802174	1	1.0
6	0.725275	0.517424	0.747475	0.821739	1	1.0
7	0.692308	0.505045	0.702128	0.825362	1	1.0
8	0.714286	0.493321	0.723404	0.824638	1	1.0
9	0.692308	0.503271	0.695652	0.821014	1	1.0
10	0.659341	0.528402	0.680412	0.786957	1	1.0
11	0.725275	0.520485	0.731183	0.802899	1	1.0
12	0.725275	0.517917	0.736842	0.811836	1	1.0
13	0.725275	0.515573	0.742268	0.811836	1	1.0
6(best according to accuracy)	0.725275	0.517424	0.747475	0.821739	1	1.0

Number of Features for Best ROC-AUC: 7

Best Features: ['original_shape_LeastAxisLength', 'original_shape_Maximum2DDiameterRow', 'original_shape_MeshVolume', 'original_shape_MinorAxisLength', 'original_shape_Sphericity', 'original_shape_SurfaceVolumeRatio', 'original_shape_VoxelVolume']

Table 5 Best features extraction RF

Number of features(KNN)	Accuracy	Log Loss	F1 Score	ROC-AUC	Precision	Recall
5	0.593407	0.929102	0.678261	0.654589	1	1.0
6	0.593407	0.929100	0.678261	0.654589	1	1.0
7	0.593407	1.009614	0.678261	0.652657	1	1.0
8	0.593407	1.114774	0.678261	0.655556	1	1.0
9	0.593407	1.231673	0.678261	0.642029	1	1.0
10	0.593407	1.231730	0.678261	0.641546	1	1.0
11	0.593407	1.311586	0.678261	0.630435	1	1.0
12	0.593407	1.397160	0.678261	0.616908	1	1.0
13	0.593407	1.435119	0.678261	0.614976	1	1.0
5(best according to accuracy)	**0.593407**	**0.929102**	**0.678261**	**0.654589**	1	1.0

Number of Features for Best ROC-AUC: 5

Best Features:['original_shape_Maximum2DDiameterRow', 'original_shape_MeshVolume', 'original_shape_MinorAxisLength', 'original_shape_SurfaceVolumeRatio', 'original_shape_VoxelVolume'],

Table 5 Best features extraction KNN

Index	Training Set	Mean	Standard Deviation
1	original_shape_Elongation	0.869154	0.074195
2	original_shape_LeastAxisLength	35.57895	5.816301
3	original_shape_MajorAxisLength	48.97369	7.160993
4	original_shape_Maximum2DDiameterColumn	54.7593	6.953782
5	original_shape_Maximum2DDiameterRow	54.47052	9.90691
6	original_shape_Maximum2DDiameterSlice	54.25526	7.304402
7	original_shape_Maximum3DDiameter	57.9414	8.835151
8	original_shape_MeshVolume	55172.9	25939.81
9	original_shape_MinorAxisLength	42.48721	6.729291
10	original_shape_Sphericity	0.764589	0.024109
11	original_shape_SurfaceArea	8973.453	2729.259
12	original_shape_SurfaceVolumeRatio	0.173649	0.025444
13	original_shape_VoxelVolume	55223.36	25947.82

Scenario2 Mean and Standard deviation for Train Set

Index	Testing Set	Mean	Standard Deviation
1	original_shape_Elongation	0.877469	0.074791
2	original_shape_LeastAxisLength	34.88897	6.012684
3	original_shape_MajorAxisLength	48.33067	7.973396
4	original_shape_Maximum2DDiameterColumn	54.33564	7.770286
5	original_shape_Maximum2DDiameterRow	53.62102	10.10456
6	original_shape_Maximum2DDiameterSlice	53.77485	7.799939
7	original_shape_Maximum3DDiameter	57.21623	9.501142
8	original_shape_MeshVolume	53684.39	28469.89
9	original_shape_MinorAxisLength	42.18842	6.604931
10	original_shape_Sphericity	0.76465	0.021585
11	original_shape_SurfaceArea	8773.71	2900.659
12	original_shape_SurfaceVolumeRatio	0.176152	0.026612
13	original_shape_VoxelVolume	53734.02	28478.28

Scenario2 Mean and Standard deviation for Test Set

	Model	Accuracy_Score	Precision_Score	Recall_Score	F1_Score
5	GBT	0.758242	0.732143	0.854167	0.755142
1	RF	0.736264	0.714286	0.833333	0.732883
2	DT	0.725275	0.709091	0.812500	0.722439
3	SVM	0.725275	0.694915	0.854167	0.719250
0	LR	0.703297	0.666667	0.875000	0.692002

Table MODEL COMPARISON for *Random undersampling*

Metric	LR	SVM	Naive Bayes classification	DT	RF	K-NN
Accuracy:	0.549	0.494	0.582	0.901	0.890	0.626
Precision	0.573	0.534	0.596	0.872	0.916	0.642
Recall	0.75	0.62	0.74	0.96	0.88	0.72
F1 Score	0.630	0.574	0.660	0.914	0.897	0.679
Log Loss	16.239	0.697	1.250	3.564	-1	13.466
Gini Coefficient	0.0658	0.038	0.130	0.789	0.782	0.232
ROC area	0.58	0.56	0.66	0.89	0.96	0.63

Figure 25 Model comparison SVMSMOTE Sampling

Cross validation	Accuracy	Log Loss	Precision	Recall	ROC AUC	F1 Score	Gini
Fold1	0.89	0.30	0.95	0.83	0.95	0.88	0.91
Fold2	0.79	0.37	0.78	0.86	0.93	0.82	0.85
Fold3	0.88	0.30	0.83	0.96	0.96	0.89	0.92
Fold4	0.89	0.30	0.94	0.86	0.95	0.90	0.91
Fold5	0.94	0.29	0.93	0.95	0.97	0.94	0.95
Average	0.88	0.31	0.89	0.89	0.95	0.89	0.91

Table Cross validation Scenario2

Appendix C:

List of the extracted Py-Radiomics Features

diagnostics_Versions_PyRadiomics: v3.1.0

diagnostics_Versions_Numpy: 1.21.6

diagnostics_Versions_SimpleITK: 2.2.1

diagnostics_Versions_PyWavelet: 1.3.0

diagnostics_Versions_Python: 3.7.0

diagnostics_Configuration_Settings: {'minimumROIDimensions': 2, 'minimumROISize': None, 'normalize': False, 'normalizeScale': 1, 'removeOutliers': None, 'resampledPixelSpacing': None, 'interpolator': 'sitkBSpline', 'preCrop': False, 'padDistance': 5, 'distances': [1], 'force2D': False, 'force2Ddimension': 0, 'resegmentRange': None, 'label': 1, 'additionalInfo': True, 'binWidth': 25, 'enableCExtensions': True}

diagnostics_Configuration_EnabledImageTypes: {'Original': {}}

diagnostics_Image-original_Hash: fdc517a332c4cb8b007f8401819250277fa36945

diagnostics_Image-original_Dimensionality: 3D

diagnostics_Image-original_Spacing: (0.30000001192092896, 0.30000001192092896, 3.5999999046325684)

diagnostics_Image-original_Size: (640, 640, 19)

diagnostics_Image-original_Mean: 0.014705489309210526

diagnostics_Image-original_Minimum: 0.0

diagnostics_Image-original_Maximum: 1.0

diagnostics_Mask-original_Hash: a4b1ab8b103a6f80c85641bea7ef997f11cfd3cc

diagnostics_Mask-original_Spacing: (0.30000001192093, 0.30000001192093, 3.5999999046325684)

diagnostics_Mask-original_Size: (640, 640, 19)

diagnostics_Mask-original_BoundingBox: (1, 1, 0, 638, 637, 19)

diagnostics_Mask-original_VoxelNum: 1777

diagnostics_Mask-original_VolumeNum: 1330

diagnostics_Mask-original_CenterOfMassIndex: (290.1789532920653, 324.2104670793472, 8.489026449071469)

diagnostics_Mask-original_CenterOfMass: (0.8694700859947062, 0.27552186732745554, -3.0028008771197037)

original_shape_Elongation: 0.7041086387017377

original_shape_Flatness: 0.29880724143804394

original_shape_LeastAxisLength: 75.73969212717414

original_shape_MajorAxisLength: 253.4734157133148

original_shape_Maximum2DDiameterColumn: 198.66525257438073

original_shape_Maximum2DDiameterRow: 149.74328626121215

original_shape_Maximum2DDiameterSlice: 230.71249257506915

original_shape_Maximum3DDiameter: 243.32830572916362

original_shape_MeshVolume: 76.9095040741324

original_shape_MinorAxisLength: 178.47282168498174

original_shape_Sphericity: 0.029925404132764715

original_shape_SurfaceArea: 2922.56452018285

original_shape_SurfaceVolumeRatio: 38.000043757476526

original_shape_VoxelVolume: 575.7480305042263

original_firstorder_10Percentile: 0.0

original_firstorder_90Percentile: 0.0

original_firstorder_Energy: 0.0

original_firstorder_Entropy: -3.203426503814917e-16

original_firstorder_InterquartileRange: 0.0

original_firstorder_Kurtosis: 0.0

original_firstorder_Maximum: 0.0

original_firstorder_MeanAbsoluteDeviation: 0.0

original_firstorder_Mean: 0.0

original_firstorder_Median: 0.0

original_firstorder_Minimum: 0.0

original_firstorder_Range: 0.0

original_firstorder_RobustMeanAbsoluteDeviation: 0.0

original_firstorder_RootMeanSquared: 0.0

original_firstorder_Skewness: 0.0

original_firstorder_TotalEnergy: 0.0

original_firstorder_Uniformity: 1.0

original_firstorder_Variance: 0.0

original_glcm_Autocorrelation: 1.0

original_glcm_ClusterProminence: 0.0

original_glcm_ClusterShade: 0.0

original_glcm_ClusterTendency: 0.0

original_glcm_Contrast: 0.0

original_glcm_Correlation: 1.0

original_glcm_DifferenceAverage: 0.0

original_glcm_DifferenceEntropy: -3.2034265038149176e-16

original_glcm_DifferenceVariance: 0.0

original_glcm_Id: 1.0

original_glcm_Idm: 1.0

original_glcm_Idmn: 1.0

original_glcm_Idn: 1.0

original_glcm_Imc1: 0.0

original_glcm_Imc2: 0.0

original_glcm_InverseVariance: 0.0

original_glcm_JointAverage: 1.0

original_glcm_JointEnergy: 1.0

original_glcm_JointEntropy: -3.2034265038149176e-16

original_glcm_MCC: 1

original_glcm_MaximumProbability: 1.0

original_glcm_SumAverage: 2.0

original_glcm_SumEntropy: -3.2034265038149176e-16

original_glcm_SumSquares: 0.0

original_gldm_DependenceEntropy: 1.5392329562896137

original_gldm_DependenceNonUniformity: 774.1660101294316

original_gldm_DependenceNonUniformityNormalized: 0.4356589815022125

original_gldm_DependenceVariance: 0.8730451536531475

original_gldm_GrayLevelNonUniformity: 1777.0

original_gldm_GrayLevelVariance: 0.0

original_gldm_HighGrayLevelEmphasis: 1.0

original_gldm_LargeDependenceEmphasis: 3.4760832864378166

original_gldm_LargeDependenceHighGrayLevelEmphasis: 3.4760832864378166

original_gldm_LargeDependenceLowGrayLevelEmphasis: 3.4760832864378166

original_gldm_LowGrayLevelEmphasis: 1.0

original_gldm_SmallDependenceEmphasis: 0.6757155330712289

original_gldm_SmallDependenceHighGrayLevelEmphasis: 0.6757155330712289

original_gldm_SmallDependenceLowGrayLevelEmphasis: 0.6757155330712289

original_glrlm_GrayLevelNonUniformity: 1735.076923076923

original_glrlm_GrayLevelNonUniformityNormalized: 1.0

original_glrlm_GrayLevelVariance: 0.0

original_glrlm_HighGrayLevelRunEmphasis: 1.0

original_glrlm_LongRunEmphasis: 1.0815274161898776

original_glrlm_LongRunHighGrayLevelEmphasis: 1.0815274161898776

original_glrlm_LongRunLowGrayLevelEmphasis: 1.0815274161898776

original_glrlm_LowGrayLevelRunEmphasis: 1.0

original_glrlm_RunEntropy: 0.14890194020807546

original_glrlm_RunLengthNonUniformity: 1663.0140642551316

original_glrlm_RunLengthNonUniformityNormalized: 0.957345562614263

original_glrlm_RunPercentage: 0.9764079477078915

original_glrlm_RunVariance: 0.03044355559584275

original_glrlm_ShortRunEmphasis: 0.9829923851148284

original_glrlm_ShortRunHighGrayLevelEmphasis: 0.9829923851148284

original_glrlm_ShortRunLowGrayLevelEmphasis: 0.9829923851148284

original_glszm_GrayLevelNonUniformity: 1330.0

original_glszm_GrayLevelNonUniformityNormalized: 1.0

original_glszm_GrayLevelVariance: 0.0

original_glszm_HighGrayLevelZoneEmphasis: 1.0

original_glszm_LargeAreaEmphasis: 2.7225563909774437

original_glszm_LargeAreaHighGrayLevelEmphasis: 2.7225563909774437

original_glszm_LargeAreaLowGrayLevelEmphasis: 2.7225563909774437

original_glszm_LowGrayLevelZoneEmphasis: 1.0

original_glszm_SizeZoneNonUniformity: 873.3127819548872

original_glszm_SizeZoneNonUniformityNormalized: 0.6566261518457799

original_glszm_SmallAreaEmphasis: 0.8373074353732025

original_glszm_SmallAreaHighGrayLevelEmphasis: 0.8373074353732025

original_glszm_SmallAreaLowGrayLevelEmphasis: 0.8373074353732025

original_glszm_ZoneEntropy: 1.0301077546727355

original_glszm_ZonePercentage: 0.7484524479459763

original_glszm_ZoneVariance: 0.9374193001300242

original_ngtdm_Busyness: 0.0

original_ngtdm_Coarseness: 1000000.0

original_ngtdm_Complexity: 0.0

original_ngtdm_Contrast: 0.0

original_ngtdm_Strength: 0.0

Bibliography

[1] Cuocolo, R., Stanzione, A., Ponsiglione, A., Romeo, V., Verde, F., Creta, M., La Rocca, R., Longo, N., Pace, L., & Imbriaco, M. (Year). Clinically significant prostate cancer detection on MRI: A radiomic shape features study. *European Journal of Radiology*. https://www.elsevier.com/locate/ejrad

[2] Shankar, V., Yang, X., Krishna, V., Tan, B., Silva, O., Rojansky, R., Ng, A., Valvert, F., Briercheck, E., Weinstock, D., Natkunam, Y., Fernandez-Pol, S., & Rajpurkar, P. (2023). LymphoML: An interpretable artificial intelligence-based method identifies morphologic features that correlate with lymphoma subtype. In Proceedings of Machine Learning Research (Vol. 225, pp. 1–17). Machine Learning for Health (ML4H) 2023.

[3] Ning, J., Li, C., Yu, P., Cui, J., Xu, X., Jia, Y., Zuo, P., Tian, J., Kenner, L., & Xu, B. (2023). Radiomic analysis will add differential diagnostic value of benign and malignant pulmonary nodules: A hybrid imaging study based on [18F]FDG and [18F]FLT PET/CT. *Insights into Imaging, 14*(197). https://doi.org/10.1186/s13244-023-01530-6

[4] Stoyanova, R., Takhar, M., Tschudi, Y., Ford, J. C., Solórzano, G., Erho, N., Balagurunathan, Y., Punnen, S., Davicioni, E., Gillies, R. J., & Pollack, A. (2016). Prostate cancer radiomics and the promise of radiogenomics. *Translational Cancer Research, 5*(4), 432–447. https://doi.org/10.21037/tcr.2016.06.20

[5] De Rooij, M., Hamoen, E. H., Fütterer, J. J., Barentsz, J. O., & Rovers, M. M. (2014). Accuracy of multiparametric MRI for prostate cancer detection: a meta-analysis. American Journal of Roentgenology, 202(2), 343-351.

[6] Steinbuss, G., Kriegsmann, M., Zgorzelski, C., Kriegsmann, K., et al. (2021). Deep learning for the classification of Non-Hodgkin lymphoma on histopathological images. *Cancers, 13*(10), Article 2419. https://doi.org/10.3390/cancers13102419

[7] Krajnc, D., Spielvogel, C. P., Grahovac, M., Ecsedi, B., Rasul, S., Poetsch, N., Traub-Weidinger, T., Haug, A. R., Ritter, Z., Alizadeh, H., ... Papp, L. (2022). Automated data preparation for in vivo tumor characterization with machine learning. Frontiers in Oncology, 12, 1017911. https://doi.org/10.3389/fonc.2022.1017911

[8] Cancer Registry of Norway. (2023). Retrieved from https://www.kreftregisteret.no/en/

[9] Fernandez-Quilez, A., Nordström, T., Jäderling, F., Kjosavik, S. R., & Eklund, M. (2023). Prostate age gap: An MRI surrogate marker of aging for prostate cancer detection. *Journal of Magnetic Resonance Imaging*.

[10] American Cancer Society. (2023). Cancer Facts & Figures 2023. Retrieved, from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html

[11] Ørsted, D. D., Nordestgaard, B. G., Jensen, G. B., Schnohr, P., & Bojesen, S. E. (2012). Prostate-specific antigen and long-term prediction of prostate cancer incidence and mortality in the general population. *European Urology, 61*(5), 865-874.

[12] TeachMeAnatomy. (2023), from https://teachmeanatomy.info/

[13] Ali, A., Feu, A., Oliveira, P., Choudhury, A., Bristow, R., & Baena, E. (2021). Prostate zones and cancer: Lost in transition?. *Nature Reviews Urology, 19*, 1-15. https://doi.org/10.1038/s41585-021-00524-7

[14] Van Booven, D. J., Kuchakulla, M., Pai, R., Frech, F. S., Ramasahayam, R., Reddy, P., Parmar, M., Ramasamy, R., & Arora, H. (2021). A Systematic Review of Artificial Intelligence in Prostate Cancer. *Research and reports in urology*, *13*, 31–39. https://doi.org/10.2147/RRU.S268596

[15] Alarcón-Zendejas, A. P., Scavuzzo, A., Jiménez-Ríos, M. A., Álvarez-Gómez, R. M., Montiel-Manríquez, R., Castro-Hernández, C., Jiménez-Dávila, M. A., Pérez-Montiel, D., González-Barrios, R., Jiménez-Trejo, F., Arriaga-Canon, C., & Herrera, L. A. (2022). The promising role of new molecular biomarkers in prostate cancer: From coding and non-coding genes to artificial intelligence approaches. *Prostate Cancer and Prostatic Diseases, 25*(3), 431-443. https://doi.org/10.1038/s41391-022-00537-2

[16] Litwin, M. S., & Tan, H. J. (2017). The Diagnosis and Treatment of Prostate Cancer: A Review. JAMA, 317(24), 2532–2542. https://doi.org/10.1001/jama.2017.7248

[17] Moser, E., Stadlbauer, A., & Windisch, C. (2009). Magnetic resonance imaging methodology. European Journal of Nuclear Medicine and Molecular Imaging, 36(Suppl 1), S30–S41. https://doi.org/10.1007/s00259-008-0938-3

[18] Bell, N., Connor Gorber, S., Shane, A., Joffres, M., Singh, H., Dickinson, J., Shaw, E., Dunfield, L., Tonelli, M., & Canadian Task Force on Preventive Health Care. (2014). Recommendations on screening for prostate cancer with the prostate-specific antigen test. CMAJ: Canadian Medical Association Journal, 186(16), 1225-1234. https://doi.org/10.1503/cmaj.140703

[19] National Institutes of Health. (2023). National Cancer Institute (NCI), from https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-cancer-institute-nci

[20] Humphrey, P. A. (2004). Gleason grading and prognostic factors in carcinoma of the prostate. *Modern Pathology, 17*(3), 292-306. https://doi.org/10.1038/modpathol.3800054

[21] Dale, B. M., Brown, M. A., & Semelka, R. C. (2015). *MRI: Basic principles and applications* (5th ed.). John Wiley & Sons, Ltd.

[22] Magnetic resonance (NMR) tomography of the brain: A preliminary clinical assessment with demonstration of pathology. Journal of Computer Assisted Tomography, 4(5), 577-586. https://doi.org/10.1097/00004728-198010000-00001

[23] Smith, F. W., Kent, C., Abramovich, D. R., & Sutherland, H. W. (1985). Nuclear magnetic resonance imaging—a new look at the fetus. *BJOG: An International Journal of Obstetrics & Gynaecology*, *92*(10), 1024-1033.

[24] Clinical functional magnetic resonance imaging service. Neurosurgery Clinics of North America, 22(2), 307-314, x. https://doi.org/10.1016/j.nec.2011.01.001 Isaacs

[25] Isaacs, B. R., Mulder, M. J., Groot, J. M., van Berendonk, N., Lute, N., Bazin, P. L., Forstmann, B. U., & Alkemade, A. (2020). 3 versus 7 Tesla magnetic resonance imaging for parcellations of subcortical brain structures in clinical settings. PLoS ONE, 15(11), Article e0236208. https://doi.org/10.1371/journal.pone.0236208

[26] Kagawa, T., Yoshida, S., Shiraishi, T., Hashimoto, M., Inadomi, D., Sato, M., Tsuzuki, T., Miwa, K., & Yuasa, K. (2017). Basic principles of magnetic resonance imaging for beginner oral and maxillofacial radiologists. Oral Radiology, 33(2), 92-100. https://doi.org/10.1007/s11282-017-0274-z

[27] Haider, M. A., van der Kwast, T. H., Tanguay, J., Evans, A. J., Hashmi, A.-T., Lockwood, G., & Trachtenberg, J. (2007). Combined T2-weighted and diffusion-weighted MRI for localization

of prostate cancer. American Journal of Roentgenology, 189(2), 323-328. https://doi.org/10.2214/AJR.06.0665

[28] Baliyan, V., Das, C. J., Sharma, R., & Gupta, A. K. (2016). Diffusion weighted imaging: Technique and applications. World Journal of Radiology, 8(9), 785-798. https://doi.org/10.4329/wjr.v8.i9.785

[29] Chen, Y., Almarzouqi, S. J., Morgan, M. L., & Lee, A. G. (2018). T2-Weighted Image. In U. Schmidt-Erfurth & T. Kohnen (Eds.), Encyclopedia of Ophthalmology (pp. 1750-1752). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69000-9_1229

[30] Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563-577. https://doi.org/10.1148/radiol.2015151169

[31] Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., & Cook, G. (2020). Introduction to radiomics. Journal of Nuclear Medicine, 61(4), 488-495. https://doi.org/10.2967/jnumed.118.222893

[32] Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging, 11, Article 91. https://doi.org/10.1186/s13244-020-00887-2

[33] Pyradiomics Documentation. https://pyradiomics.readthedocs.io/en/latest/

[34] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2, Article 160. https://doi.org/10.1007/s42979-021-00592-x

[35] Elton, E. J., Gruber, M. J., & Busse, J. A. (1998). Do investors care about sentiment?. *The Journal of Business*, *71*(4), 477-500.

[36] Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, *316*(5), 533-534.

[37] Schwartzman, O., Savino, A. M., Gombert, M., Palmi, C., Cario, G., Schrappe, M., ... & Izraeli, S. (2017). Suppressors and activators of JAK-STAT signaling at diagnosis and relapse of acute lymphoblastic leukemia in Down syndrome. *Proceedings of the National Academy of Sciences*, *114*(20), E4030-E4039.

[38] Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., ... & Bischl, B. (2023, July). Relating the partial dependence plot and permutation feature importance to the data generating process. In *World Conference on Explainable Artificial Intelligence* (pp. 456-479). Cham: Springer Nature Switzerland.

[39] Baturynska, I., & Martinsen, K. (2021). Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. *Journal of Intelligent Manufacturing*, *32*, 179-200.

[40] sklearn.manifold.TSNE. scikit-learn developers, 2021, https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (accessed 22.11.2021)

[41] Undersampling Algorithms for Imbalanced Classification. Jason Brownlee, 2020, https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/ (accessed 23.11.2021)

[42] Validation curves: plotting scores to evaluate models. https://scikit-learn.org/stable/modules/learning_curve.html

[43] Githib link: https://github.com/RihabHashim/Master-Project

[44] Ting, K. M. (2010). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 209-209). Springer US. https://doi.org/10.1007/978-0-387-30164-8_157