ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE E TECNOLOGIE DELLA SALUTE

Ciclo 37

Settore Concorsuale: 06/N1 - SCIENZE DELLE PROFESSIONI SANITARIE E DELLE TECNOLOGIE MEDICHE APPLICATE

Settore Scientifico Disciplinare: MED/50 - SCIENZE TECNICHE MEDICHE APPLICATE

APPLICATION AND DEVELOPMENT OF AN ARTIFICIAL INTELLIGENCE BASED MULTIPARAMETRIC MALIGNANCY INDEX (MMI) FOR INTERPRETATION OF 3T MULTIPARAMETRIC PROSTATE MRI

Presentata da: Fabio Ferroni

Coordinatore Dottorato

Igor Diemberger

Supervisore

Alessandro Bevilacqua

Co-supervisore

Domenico Barone

Esame finale anno 2025

Abstract

Introduction: Prostate cancer (PCa) diagnosis and management remain challenging due to the disease's varied progression, which ranges from slow-growing, indolent forms to aggressive types needing early intervention. While recent advancements in detection, including risk-adapted screening and multiparametric MRI (mpMRI), have improved early diagnosis, balancing the identification of high-risk cases while minimizing overdiagnosis remains a critical research focus. This study explores the application of machine learning (ML), specifically a Random Forest (RF) model, to enhance PCa detection accuracy by classifying cases based on clinical and radiological features.

Materials and Methods: A cohort of 314 patients aged 55-75 years was analyzed using key variables, including PSA levels, PI-RADS scores, prostate volume, and rADC. The RF model was selected for its robustness in handling mixed data types and complex feature interactions. To train and evaluate the model, patients were stratified into ISUP grades (0, 1, 2+), with non-cancerous (ISUP 0) and aggressive (ISUP 2+) cases distinctly represented. A subset of 95 patients was reserved for testing to assess the model's generalization capabilities.

Results: Within the cohort, 182 cases were classified with lower PI-RADS scores (2 and 3), and 132 were classified with higher scores (4 and 5). The RF model achieved an accuracy of 68% on the test set, with precision and recall metrics for each ISUP grade: ISUP 0 (precision: 0.79, recall: 0.83), ISUP 1 (precision: 0.22, recall: 0.12), and ISUP 2+ (precision: 0.58, recall: 0.74). Feature importance analysis highlighted rADC, prostate volume, and PSA as top predictors, with an AUC score further underscoring the model’s ability to distinguish between non-cancerous and aggressive PCa cases.

Conclusions: This study supports the potential of ML in improving diagnostic accuracy for PCa. The RF model, with its effective handling of clinical and radiological features, provides insights into significant predictors that could aid clinicians in risk stratification. Additionally, this work aligns with the aims of the FLUTE project, a European initiative seeking to leverage federated learning for enhanced PCa diagnostics across institutions. Although FLUTE is currently paused, the groundwork laid through this thesis offers a foundation for future AI-driven diagnostic tools, moving toward more accurate and personalized treatment planning in PCa management.

“We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists”.

Geoffrey Hinton, AI pioneer and neural network innovator, 2016

List of Abbreviations

ADC - Apparent Diffusion Coefficient

ADT - Androgen Deprivation Therapy

AI - Artificial Intelligence

AUC - Area Under the Curve

AS - Anterior Stroma (prostate region)

BPH - Benign Prostatic Hyperplasia

BRCA - Breast Cancer gene (linked to hereditary cancer risks)

csPCa - Clinically Significant Prostate Cancer

CZ - Central Zone of the prostate

DCE - Dynamic Contrast-Enhanced Imaging

DPIA - Data Protection Impact Assessment

DRE - Digital Rectal Examination

DWI - Diffusion-Weighted Imaging

FL - Federated Learning

FLUTE - Federated Learning and Multi-party Computation Techniques for Prostate Cancer

FN - False Negative

FP - False Positive

FPR - False Positive Rate

FAIR - Findability, Accessibility, Interoperability, and Reusability

GS - Gleason Score

HIFU - High-Intensity Focused Ultrasound

ISUP - International Society of Urological Pathology

KNN - K-Nearest Neighbors (machine learning algorithm)

ML - Machine Learning

mpMRI - Multiparametric Magnetic Resonance Imaging

MR - Magnetic Resonance

MRI/US - MRI/Ultrasound Fusion Imaging

NN - Neural Network

ncsPCa - Non-Clinically Significant Prostate Cancer

PACS - Picture Archiving and Communication System

PCA - Principal Component Analysis (statistical technique)

PCa - Prostate Cancer

PET - Positron Emission Tomography

PI-RADS - Prostate Imaging Reporting and Data System

PSA - Prostate-Specific Antigen

PSAD - Prostate-Specific Antigen Density

PSMA - Prostate-Specific Membrane Antigen

PZ - Peripheral Zone of the prostate

rADC - Ratio of Apparent Diffusion Coefficient

RF - Random Forest

ROC - Receiver Operating Characteristic

RT - Radiotherapy

SHAP - SHapley Additive exPlanations (model interpretability method)

SMOTE - Synthetic Minority Over-sampling Technique

SMPC - Secure Multi-Party Computation

TEE - Trusted Execution Environment

TNM - Tumor-Node-Metastasis (cancer staging system)

TPR - True Positive Rate

TRUS - Transrectal Ultrasound

T2w - T2-Weighted Imaging

TZ - Transition Zone of the prostate

1. Introduction

The prostate is a small glandular organ located in the male pelvis, just below the bladder and anterior to the rectum. It encircles the prostatic urethra, the portion of the urethra that passes through the prostate, playing a key role in both urinary and reproductive systems.

Anatomically, the prostate is divided into distinct zones: the peripheral zone (PZ), which constitutes the largest part of the gland and is the most common site for prostate cancer (PCa), the central zone (CZ) surrounding the ejaculatory ducts, and the transition zone (TZ), which often enlarges in benign prostatic hyperplasia (BPH).

The prostate is encapsulated by a fibromuscular stroma, providing structural support, and it is closely associated with important anatomical structures like the seminal vesicles, neurovascular bundles, and the bladder neck. This intricate anatomical arrangement makes the prostate not only vital for reproductive function, by producing seminal fluid, but also a key structure in maintaining urinary continence and normal urinary flow. Understanding its anatomy is crucial for the accurate diagnosis and treatment of prostate-related diseases.

1.1. Overview of PCa

PCa is one of the most common malignancies affecting men worldwide, particularly in developed countries. PCa is a highly heterogeneous disease, characterized by variable clinical behavior ranging from slow-growing, indolent tumors to aggressive forms that can rapidly metastasize and lead to significant morbidity and mortality. The complexity of PCa arises from its diverse presentation and progression, requiring personalized approaches for its diagnosis, management, and treatment.

The clinical management of PCa has significantly evolved over the last few decades. Early detection strategies, such as prostate-specific antigen (PSA) testing, along with advancements in imaging techniques, have contributed to identifying the disease at earlier stages. However, overdiagnosis and overtreatment remain significant challenges, particularly in cases of low-risk cancer. Thus, there is a strong emphasis on risk stratification and personalized care, which includes active surveillance for low-risk patients and more aggressive interventions for high-risk and advanced-stage diseases.

1.2. Epidemiology and Risk Factors

PCa is the second most commonly diagnosed cancer in men, with a higher incidence in developed countries. The incidence increases with age, with most cases diagnosed in men over the age of 65. This makes aging one of the strongest risk factors for PCa. Additionally, ethnic background plays a significant role, as African-American men are more likely to develop PCa and have a higher risk of aggressive disease compared to Caucasians and Asians. Genetic predisposition is also a major risk factor, with mutations in the BRCA1 and BRCA2 genes, as well as a family history of PCa, contributing to an increased risk.

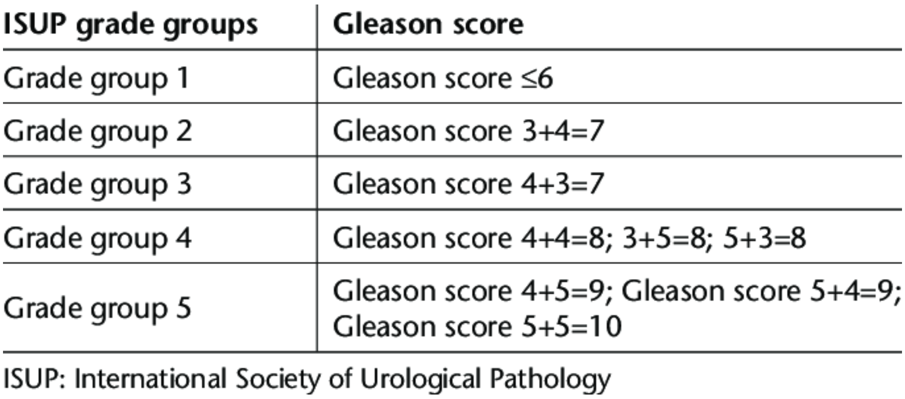
While age, ethnicity, and genetics are well-established risk factors, there is currently no strong evidence supporting specific lifestyle modifications or preventive measures to reduce the risk of

PCa. However, ongoing research continues to explore the role of diet, inflammation, and environmental factors in the development of the disease.

1.3. Pathophysiology and Classification

PCa typically originates in the glandular cells of the prostate, and it is classified as an adenocarcinoma. The disease's progression is often slow, but a subset of tumors can behave aggressively and metastasize to other organs, such as bones and lymph nodes. The Tumor-Node-Metastasis (TNM) classification system is the standard for staging PCa and is essential for guiding treatment decisions. The 2017 TNM classification used in clinical practice helps categorize the primary tumor (T), regional lymph node involvement (N), and distant metastasis (M), providing a structured framework for understanding the extent of the disease.

Histological evaluation of prostate tissue remains the gold standard for diagnosing and classifying PCa. The Gleason score (GS) and its updated version by the International Society of Urological Pathology (ISUP) are widely used to assess the aggressiveness of the tumor. The ISUP system stratifies PCa into five grade groups, based on the Gleason score, with group 1 representing low-risk tumors and group 5 encompassing high-risk, aggressive cancers. These grading and staging systems are vital for distinguishing between clinically significant PCa (csPCa), which requires intervention, and insignificant PCa (ncsPCa), which may be managed with active surveillance.



ISUP grade groups	Gleason score
Grade group 1	Gleason score ≤6
Grade group 2	Gleason score 3+4=7
Grade group 3	Gleason score 4+3=7
Grade group 4	Gleason score 4+4=8; 3+5=8; 5+3=8
Grade group 5	Gleason score 4+5=9; Gleason score 5+4=9; Gleason score 5+5=10

ISUP: International Society of Urological Pathology

1.4. Early Detection and Diagnosis

Early detection of PCa typically involves PSA testing, digital rectal examination (DRE), and biopsy. PSA may be elevated in cases of PCa, benign prostatic hyperplasia (BPH), or prostatitis. Although PSA testing has improved early detection, it is not without limitations. The risk of overdiagnosis (identifying cancers that may never progress or cause harm) has led to increased scrutiny of PSA-based screening programs.

To mitigate the risks of overdiagnosis, current guidelines advocate for a risk-adapted strategy in screening. This involves assessing individual risk factors, such as age, family history of PCa and genetic predisposition, to tailor screening efforts. For example, men with a family history of PCa or those carrying BRCA mutations may benefit from earlier and more frequent screening.

Definitive diagnosis of PCa requires histopathological verification through prostate biopsy. Advances in magnetic resonance imaging (MRI) have improved the accuracy of biopsies, particularly when using MRI-targeted biopsies. These techniques help identify clinically significant lesions, reducing the likelihood of unnecessary biopsies in men with low-risk disease.

1.5. Treatment Strategies

The management of PCa depends on the stage of the disease, the patient's life expectancy, and the potential risks and benefits of treatment. Localized disease, which is confined to the prostate, may be managed through active surveillance, surgery (radical prostatectomy), or radiation therapy. Active surveillance is often recommended for men with low-risk, localized disease, allowing for regular monitoring without immediate treatment. This approach minimizes the risk of overtreatment and its associated side effects, such as urinary incontinence and sexual dysfunction.

For intermediate- and high-risk PCa, curative treatment options include radical prostatectomy or radiation therapy, often combined with androgen deprivation therapy (ADT). ADT is designed to lower androgen levels or block androgen receptor signaling, thereby slowing the growth of cancer. In cases where the disease has metastasized, systemic therapies, including chemotherapy and next-generation androgen receptor inhibitors like abiraterone and enzalutamide, are recommended to prolong survival and improve quality of life.

Emerging treatments for advanced PCa, such as prostate-specific membrane antigen (PSMA) PET/CT imaging and radioligand therapies, offer more targeted approaches, particularly for men with metastatic castration-resistant PCa. These innovations represent significant advancements in the management of advanced disease, providing more effective treatment options with potentially fewer side effects.

In conclusion, PCa remains a complex and multifaceted disease, requiring a tailored approach to its diagnosis and management. Advances in early detection, particularly risk-adapted screening strategies and improvements in imaging, have enhanced the ability to identify PCa at earlier, more treatable stages. The challenge of balancing the benefits of early detection with the risks of overdiagnosis and overtreatment remains a critical focus of current research.

AI models show great promise in addressing these challenges by supporting more accurate and standardized diagnostic processes. In line with this potential, tools like chatgpt contributed to the refinement and clarity of this thesis. PCa continues to pose a significant challenge in clinical practice due to its highly variable nature, with progression ranging from slow-growing, indolent forms to aggressive disease requiring timely intervention. Accurate early detection and precise characterization are key to ensuring appropriate treatment strategies, which is where multiparametric magnetic resonance imaging (mpMRI) has emerged as a crucial tool. MpMRI enables the detailed visualization of the prostate, offering valuable insights into both anatomical and functional changes associated with malignancy. Despite these advances, the variability in interpreting mpMRI results remains a challenge, underscoring the need for improved standardization and diagnostic accuracy [1].

2. The Role of Magnetic Resonance Imaging in PCa Diagnosis

MpMRI has emerged as an essential tool in the diagnostic pathway for PCa. This chapter discusses the role of MRI in the detection, characterization, and risk stratification of PCa, providing an evidence-based overview of its clinical utility, limitations, and future directions.

2.1. Role of MRI in PCa Detection

Historically, PCa diagnosis relied heavily on PSA levels, DRE, and systematic transrectal ultrasound (TRUS)-guided biopsies. However, these methods presented challenges, including over-detection of indolent cancers and under-detection of aggressive disease. The advent of mpMRI has significantly improved the diagnostic accuracy, particularly for detecting csPCa, while reducing unnecessary biopsies.

MpMRI combines three imaging sequences: T2-weighted imaging (T2w), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCE). Each of these sequences provides specific information that aids in distinguishing between benign conditions and malignancies. For instance, PCa typically appears as areas of low signal intensity on T2w, with restricted diffusion on DWI, and early enhancement on DCE sequences. These features enhance the detection and localization of csPCa, particularly those of ISUP grade group ≥2.

The Prostate Imaging-Reporting and Data System (PI-RADS), currently in version 2.1, provides a standardized approach to interpreting mpMRI results. The PI-RADS scoring system stratifies lesions on a scale from 1 to 5, where higher scores correspond to a greater likelihood of csPCa. This standardization has significantly reduced interobserver variability and has improved communication between radiologists and clinicians.

2.2. MRI in Biopsy Decision-Making and Targeted Biopsies

The integration of MRI into the diagnostic workflow has been a significant advancement in the indication for prostate biopsies. Historically, men with elevated PSA levels or abnormal DRE results would undergo systematic biopsies, which involve taking 12 cores from predefined regions of the prostate. Systematic biopsy has been associated with over-diagnosis of indolent cancers (e.g. ISUP grade group 1), and under-sampling of csPCa.

MRI has transformed this paradigm by enabling MRI-targeted biopsy, which can be performed using cognitive guidance, MRI/ultrasound (US) fusion software, or direct in-bore MRI-guided biopsy. These techniques have shown to significantly improve the detection rate of csPCa while reducing the detection of low-risk lesions, which may not require immediate treatment. This is particularly evident in patients with prior negative biopsies but persistent suspicion of PCa. MRI- targeted biopsy has consistently demonstrated superiority over systematic biopsy in detecting ISUP grade group ≥2 cancers.

A meta-analysis comparing MRI to template biopsies in biopsy-naive and repeat-biopsy settings reported a pooled sensitivity of 91% for ISUP grade group ≥2 cancers, compared to systematic biopsy alone. This increased accuracy minimizes unnecessary interventions and reduces the risk of overtreatment, making MRI-targeted biopsy a more precise and efficient diagnostic tool.

2.3. Screening and Risk Stratification

While MRI is not yet recommended as a primary screening tool for PCa due to cost and resource constraints, it plays a critical role following initial PSA testing. When PSA levels are elevated, MRI can help to stratify the risk of csPCa and guide the decision to perform a biopsy. By incorporating MRI findings into risk calculators, clinicians can better predict biopsy outcomes and avoid unnecessary procedures.

Recent studies, such as the Stockholm3 trial, have demonstrated that using MRI in conjunction with PSA screening significantly reduces the number of biopsies performed, while improving the detection rate of csPCa . In addition, trials such as the PRECISION and MRI-FIRST have confirmed the efficacy of MRI as a triage tool, showing that an MRI-first approach results in fewer unnecessary biopsies and improved detection of csPCa .

2.4. Limitations and Future Directions

Despite its many advantages, MRI is not without limitations. One challenge is its lower sensitivity for detecting ISUP grade group 1 cancers, which can sometimes lead to underdiagnosis in cases where the cancer is not visible on MRI. Furthermore, the interpretation of MRI can vary depending on the radiologist’s experience and the quality of the imaging equipment. This highlights the need for standardized protocols and continued training to ensure accurate and reliable results .

Additionally, while MRI-targeted biopsy reduces the detection of insignificant cancers, there remains a small risk of missing significant cancers, particularly in patients with negative MRI findings but persistent clinical suspicion. Ongoing research aims to refine MRI-based techniques and explore the use of artificial intelligence (AI) to enhance the accuracy and predictive value of MRI in PCa diagnosis .

Looking ahead, the integration of advanced imaging modalities, such as prostate-specific membrane antigen (PSMA) PET/MRI, may further enhance the ability to detect and characterize PCa, particularly in cases where MRI alone may be insufficient. These emerging technologies hold promise for improving diagnostic precision and reducing the burden of PCa.

2.5. Conclusion

MRI has become a cornerstone in the diagnostic pathway for PCa. Its ability to accurately detect and localize csPCa, while reducing unnecessary biopsies, has revolutionized PCa diagnostics. As MRI technology continues to evolve, it will likely play an even more prominent role in personalized

PCa management, guiding not only diagnosis but also treatment decisions and follow-up strategies.

In summary, the use of MRI in PCa diagnosis enhances the precision of biopsies, facilitates early detection of csPCa, and reduces the risks associated with overdiagnosis and overtreatment. Continuing advancements in imaging technology and interpretation will undoubtedly further cement MRI’s role as an indispensable tool in the fight against PCa [1].

3. The PI-RADS System: A Framework for mpMRI Assessment

3.1 Introduction to PI-RADS

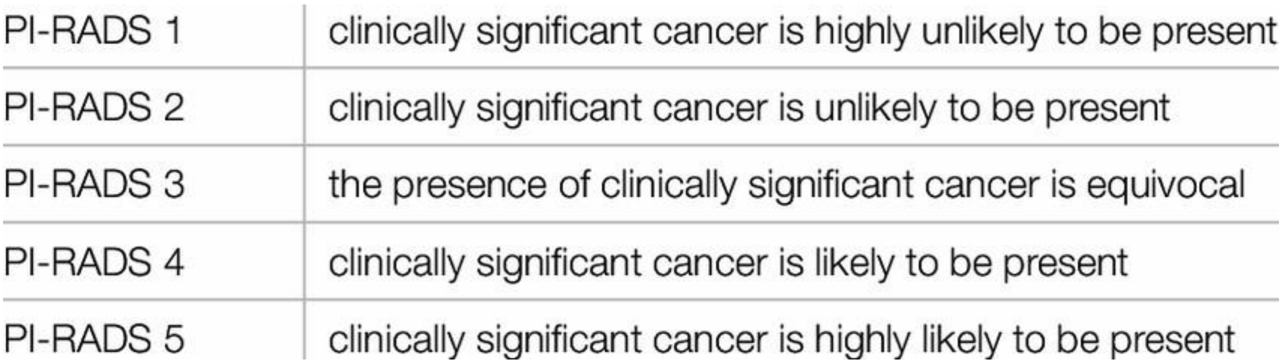
The Prostate Imaging Reporting and Data System (PI-RADS) is an internationally recognized framework designed to standardize the acquisition, interpretation, and reporting of mpMRI for PCa detection. It was initially developed in response to the need for greater consistency in prostate MRI protocols, given the variability in diagnostic accuracy across different institutions and radiologists. PI-RADS aims to improve the identification of csPCa while minimizing unnecessary biopsies and treatments for indolent cancers.

3.2 Development and Evolution of PI-RADS

PI-RADS was first introduced by the European Society of Urogenital Radiology (ESUR) in 2012, with version 1 (v1) providing the initial structured approach to prostate MRI. However, due to rapid advancements in imaging technology and increasing clinical experience, limitations of v1 became evident, particularly regarding interobserver variability and technical aspects of mpMRI. To address these issues, the American College of Radiology (ACR), ESUR, and the AdMeTech Foundation collaborated to release PI-RADS version 2 (v2) in 2015, followed by the most recent update, PI-RADS version 2.1 (v2.1), in 2019.

3.3 Core Components of PI-RADS v2.1

PI-RADS v2.1 is designed to evaluate treatment-naïve prostate glands using a standardized scoring system based on the analysis of three primary MRI sequences: T2w, DWI, and DCE. The system assigns scores ranging from 1 to 5 based on the likelihood that a particular lesion corresponds to csPCa, where:



PI-RADS 1	clinically significant cancer is highly unlikely to be present
PI-RADS 2	clinically significant cancer is unlikely to be present
PI-RADS 3	the presence of clinically significant cancer is equivocal
PI-RADS 4	clinically significant cancer is likely to be present
PI-RADS 5	clinically significant cancer is highly likely to be present

In PI-RADS v2.1, the dominant sequence for assessing the PZ is DWI, while for the TZ, T2W imaging takes precedence. DCE plays a supportive role, particularly in ambiguous cases where DWI findings are indeterminate (PI-RADS 3). This multiparametric approach enhances the diagnostic accuracy of prostate MRI by integrating both anatomical and functional data to evaluate the presence and aggressiveness of PCa.

3.4 PI-RADS Scoring and Clinical Application

The standardized scoring system is central to PI-RADS, providing a structured way to communicate imaging findings to clinicians, which informs subsequent patient management decisions, such as whether to proceed with biopsy or to pursue active surveillance. For example, lesions scored as PI-RADS 4 or 5 are highly suggestive of csPCa, often prompting a targeted biopsy. In contrast, lesions scored as PI-RADS 1 or 2 are unlikely to represent significant cancer, reducing the need for invasive procedures.

3.5 Clinical Impact and Ongoing Research

PI-RADS v2.1 has proven to be a valuable tool in improving the detection and management of PCa, significantly influencing clinical practice by enhancing diagnostic confidence and reducing unnecessary biopsies. However, ongoing research is focused on refining the system further, particularly in integrating advanced imaging techniques such as AI, MR spectroscopy and diffusion tensor imaging, which may be incorporated into future versions as clinical data accumulates [2].

4. Research project

4.1. Protocol Design and IRB approval

A comprehensive review of the existing literature on mpMRI was conducted, with a particular focus on advancements in quantitative MRI and radiomic features. Recent research has emphasized the potential of quantitative MRI metrics and radiomic analysis to enhance diagnostic accuracy and prognostic assessments, enabling more personalized treatment strategies for patients with PCa.

The current project was reviewed and approved by the Institutional Review Board (IRB) in March 2022. This approval ensures that the study adheres to ethical standards in conducting research involving human subjects, with particular attention to data privacy, and scientific rigor. The findings of this project aim to contribute to the growing body of knowledge on how mpMRI, enhanced by quantitative and radiomic approaches, can improve the diagnosis and management of PCa, ultimately leading to better patient outcomes.

4.2. Patient enrollment

For this study, we carried out an extensive retrospective review using the Picture Archiving and Communication System (PACS) of our institution, IRCCS IRST. The aim was to identify all patients who underwent 3T mpMRI over a three-year period, specifically from January 2018 to December 2020. The criteria for inclusion required patients to have undergone MRI scans following a standardized protocol, with imaging performed using a 3T multicoil Ingenia MRI system by Philips. The standardized mpMRI protocols used for these patients included multiple key sequences: T2w imaging for anatomical detail, DWI to assess the diffusion of water molecules in tissues, Apparent Diffusion Coefficient (ADC) maps derived from DWI, and DCE sequences, which are instrumental in visualizing vascular patterns and perfusion characteristics in the prostate tissue.

In addition to mpMRI, all patients underwent a transrectal ultrasound (TRUS)-guided biopsy of the prostate. This procedure was performed as part of their standard clinical care. The biopsies provided histopathological confirmation of the presence or absence of PCa for all selected patients. For patients who, as part of their therapeutic pathway, underwent radical prostatectomy, the histopathological report was carefully reviewed. This step was essential to obtain definitive confirmation of the presence and characteristics of the prostate tumor, allowing for a direct correlation between the findings obtained through mpMRI and the histopathological features observed in the surgical specimen. The analysis of the histopathological report provided crucial information on tumor staging, GS, extracapsular extension, and surgical margin involvement, among other relevant prognostic factors.

To maintain the integrity and quality of the dataset, strict exclusion criteria were applied. Patients were not included in the study if they had any condition that could interfere with the quality or interpretation of the MRI images. Specifically, individuals with a hip prosthesis, which can cause significant artifacts on pelvic imaging, were excluded. Similarly, patients with prominent hemorrhage, which might obscure or distort the appearance of the prostate, were also excluded. Further exclusions included patients who had previously undergone radiotherapy (RT) or focal therapies such as high-intensity focused ultrasound (HIFU) or cryotherapy, as these treatments can significantly alter the prostate's tissue characteristics, potentially confounding imaging interpretations. Lastly, patients whose mpMRI images were affected by severe motion artifacts were not included in the study, as these artifacts could degrade the quality of the images and reduce the reliability of the analysis.

By adhering to these rigorous inclusion and exclusion criteria, we ensured that the study population consisted of patients whose mpMRI, biopsy, and, when applicable, histopathological data from radical prostatectomy could be reliably compared. This approach allowed for a more comprehensive evaluation, supporting a more accurate assessment of the diagnostic utility of mpMRI in the detection, characterization, and correlation with the final pathological outcomes of PCa.

4.3. Manual database creation

We manually created a comprehensive database using an Excel spreadsheet, integrating clinical, imaging, and pathological data from various sources. The clinical information—such as age, PSA levels, PSA density (calculated as PSA divided by prostate volume), DRE findings, and family history of PCa—was collected through an anamnesis form that each patient completed at the time of their mpMRI examination. For the MRI data, we carefully reviewed the radiological reports, focusing on image quality and the PI-RADS score for each lesion. To maintain clarity and relevance, we summarized the MRI findings by prioritizing the PI-RADS score of each identified lesion and its anatomical localization within the prostate. This simplified approach allowed us to capture the essential imaging data without unnecessary complexity.

The pathological data, including results from both previous and subsequent prostate biopsies, as well as final anatomopathological reports for patients who underwent radical prostatectomy, were obtained from the institutional medical record system, Log80, used by AUSL Romagna. For these reports, some information was interpreted and summarized for inclusion in the database. Specifically, we extracted and recorded the highest GS and ISUP grade from either biopsy or surgical reports, ensuring that the most clinically relevant data were captured. This summarized approach provided a clear representation of each patient's pathology results.

Through the meticulous compilation and analysis of specific data points, combined with the interpretation of critical elements such as the most significant biopsy or surgical findings, we developed a comprehensive and precise dataset. This Excel database facilitated an in-depth comparison of clinical, imaging, and histopathological findings. It provided a solid foundation for evaluating the diagnostic utility of mpMRI in detecting and characterizing PCa, with definitive pathological outcomes serving as the cornerstone of our analysis.

4.4. Structured report and automatic database implementation

Introduction

In modern radiology, structured reporting has emerged as an essential tool to enhance communication between radiologists, clinicians, and researchers. The transition from traditional narrative reporting to structured formats has brought significant improvements in clarity, consistency, and usability of radiological data. This chapter explores the development and integration of a structured reporting system for mpMRI within our research institution, focusing on the automated compilation of a clinical-radiological database. We will examine the processes involved in creating this system and discuss its benefits, particularly in terms of data standardization and ease of access for future research and clinical applications.

The Role of Structured Reporting in Radiology

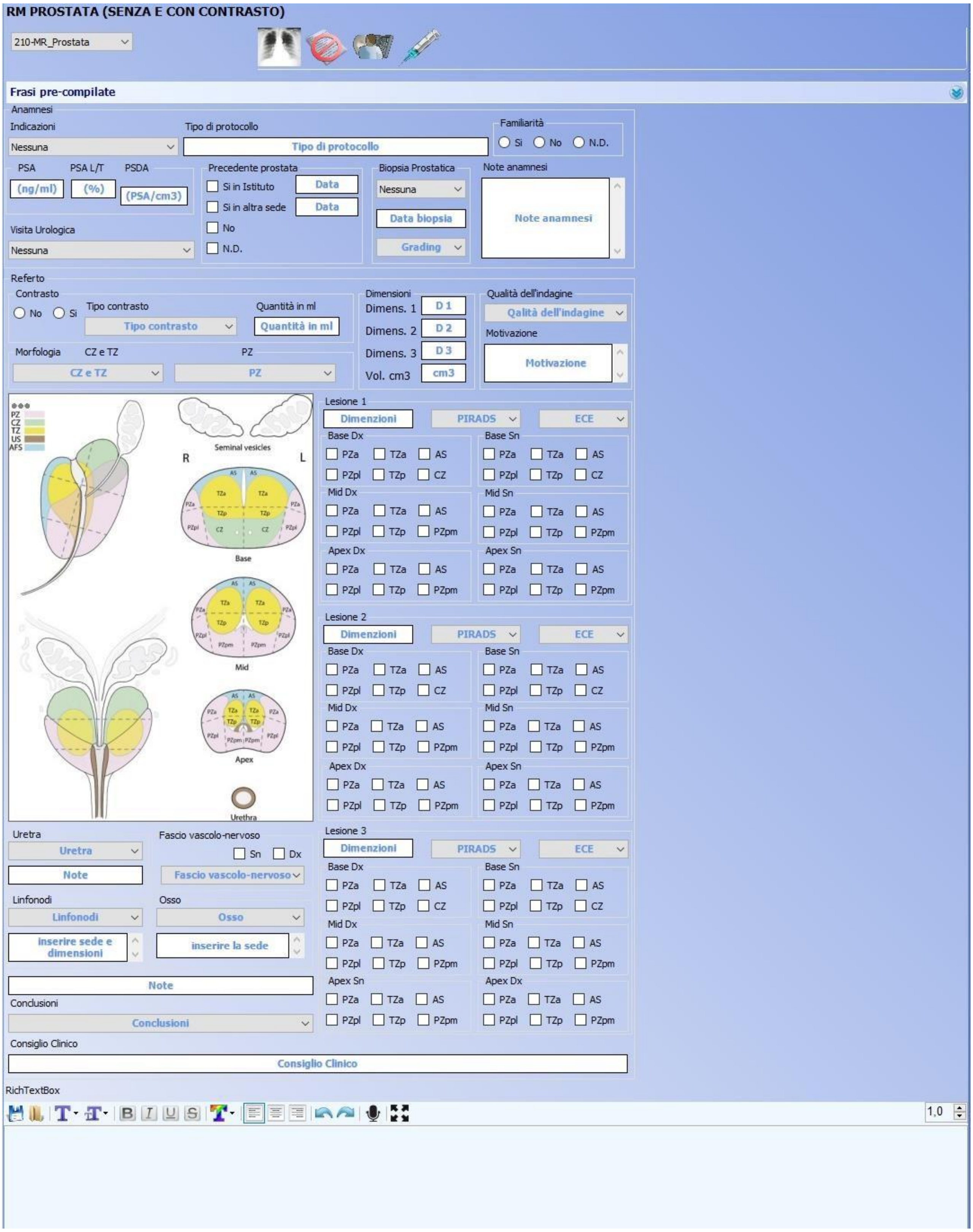
Radiological reporting has historically been narrative-based, relying on free-text descriptions that often vary in format, language, and detail. This approach, while flexible, presents challenges for data consistency, particularly when attempting to extract information for research purposes or for machine learning (ML) algorithms. Structured reporting addresses these issues by providing a standardized format that ensures essential information is consistently documented across all cases.

For instance, the PI-RADS has been developed to standardize the reporting of mpMRI for PCa detection. In our institution, we collaborated with AGFA HealthCare to create a structured report specifically tailored for mpMRI, based on the PI-RADS v2.1 system. This report enhances clarity for both patients and specialists, enabling quick and efficient interpretation of findings. It includes key clinical and radiological data, such as PI-RADS scores for lesions, anatomical localization, and image quality assessment, ensuring that all relevant information is captured in a standardized format.

Development and Integration of the Structured Report

The process of developing the structured report involved multiple stages. Initially, clinical input was provided to AGFA HealthCare to design a report that met the needs of both radiologists and referring physicians. The structured report was created with a graphical interface that allows easy data entry, focusing on standardizing the collection of clinical and imaging data. This report includes fields for clinical history (e.g., PSA levels, PSAD), DRE results, family history for PCa, and MRI findings such as PI-RADS scores and lesion localization.

To streamline data collection and minimize human error, we integrated this structured report into our institution’s Radiology Information System (RIS). This integration allowed for the automatic extraction and anonymization of clinical and radiological data directly from the patient's report. The automatic export of this data into an Excel format has significantly improved workflow efficiency, eliminating the need for manual data entry, which was previously time-consuming and prone to inconsistencies.



RM PROSTATA (SENZA E CON CONTRASTO)

210-MR_Prostata

Frasi pre-compilate

Anamnesi

Indicazioni: Nessuna

Tipo di protocollo: Tipo di protocollo

Familiarità: ○ Sì ○ No ○ N.D.

PSA (ng/ml) PSA L/T (%) PSDA (PSA/cm3)

Precedente prostata:
□ Sì in Istituto Data
□ Sì in altra sede Data
□ No
□ N.D.

Biopsia Prostatica: Nessuna
Data biopsia
Grading

Note anamnesi: Note anamnesi

Visita Urologica: Nessuna

Referto

Contrasto: ○ No ○ Sì
Tipo contrasto: Tipo contrasto
Quantità in ml: Quantità in ml

Dimensioni
Dimens. 1 D 1
Dimens. 2 D 2
Dimens. 3 D 3
Vol. cm3 cm3

Qualità dell'indagine: Qalità dell'indagine

Motivazione: Motivazione

Morfologia
CZ e TZ: CZ e TZ
PZ: PZ

Lesione 1: Dimenzioni PIRADS ECE
Base Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Base Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Mid Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Mid Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm

Lesione 2: Dimenzioni PIRADS ECE
Base Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Base Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Mid Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Mid Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm

Lesione 3: Dimenzioni PIRADS ECE
Base Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Base Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ CZ
Mid Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Mid Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Sn: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm
Apex Dx: □ PZa □ TZa □ AS □ PZpl □ TZp □ PZpm

Uretra: Uretra
Note

Fascio vascolo-nervoso: □ Sn □ Dx
Fascio vascolo-nervoso

Linfonodi: Linfonodi
inserire sede e dimensioni

Osso: Osso
inserire la sede

Note

Conclusioni: Conclusioni

Consiglio Clinico: Consiglio Clinico

RichTextBox

1,0

Automation of Clinical-Radiological Database Creation

Before the implementation of structured reporting, we manually compiled databases of patients undergoing mpMRI. This task involved collecting data from multiple sources, including patient anamnesis forms and radiological reports, which had to be transcribed and organized manually.

Such manual processes were labor-intensive, time-consuming, and increased the risk of data entry errors.

To address these challenges, we implemented an automated system that compiles data from both the structured radiology reports and the clinical information stored in the hospital's repository (Log80), used by AUSL Romagna.

This system automatically retrieves clinical and radiological data, has the capability to anonymize it, and compiles it into a single database. The automation process involves the extraction of key variables such as age, PSA levels, PSAD, as well as the PI-RADS score and anatomical localization of lesions.

Additionally, this procedure allows for direct access to the anatomopathological report or biopsy results without the need to search for them in Log80. However, the system still requires the physician to read, summarize, and interpret the report, assigning the ISUP grade and entering it into the database. While the system does not yet provide a fully automated process, it significantly facilitates and reduces the amount of manual work involved, improving overall workflow efficiency.

The system is expected to be further developed and enhanced over time, with the goal of eventually enabling automatic updates to the database as new patient data becomes available. This future implementation will not only streamline the process further but also ensure that the database remains current and reflective of the most up-to-date clinical and pathological information for each patient.

Benefits of Structured Reporting and Automation

The integration of structured reporting and automated database compilation has provided several key benefits to our institution:

- Improved Data Standardization: Structured reporting ensures that essential clinical and imaging data are consistently documented across all cases, reducing variability and enhancing the reliability of the data.
- Increased Efficiency: Automation of data extraction and anonymization has drastically reduced the time required to compile and manage clinical-radiological databases, allowing us to focus on more complex tasks such as data analysis and research.
- Enhanced Research Capabilities: The availability of standardized, high-quality data facilitates the use of big data techniques and ML in radiology. The structured reports serve as a reliable source of ground truth for training AI algorithms aimed at improving PCa diagnosis and management .
- Clinical Benefits: For clinicians, structured reports provide a clear, concise summary of patient data, aiding in decision-making and treatment planning. Additionally, patients benefit from reports that present information in an easily understandable format, improving communication and patient engagement.

Conclusion

The transition to structured reporting and automated data compilation has significantly improved the workflow within our radiology department, providing a more efficient and reliable means of managing clinical-radiological data. By integrating these systems, we have not only streamlined the process of database creation but also enhanced the quality and consistency of the data available for research and clinical applications. This innovation underscores the importance of structured reporting in modern radiology, particularly in the context of big data and AI, where accurate, standardized data is crucial for future advancements.

Moreover, this initiative was driven by the fact that, during the course of my PhD, both my institution and I were awarded a European research grant through the FLUTE project. This funding created the necessity to implement and expand our manually curated databases, ensuring they met the high standards required for large-scale research. The transition to an automated and structured system was essential to meet the project's objectives, enabling us to handle the increasing amount of data efficiently and support the research efforts tied to the FLUTE project [3].

4.5. FLUTE project (Federated Learning and mUlti-party computation Techniques for prostatE cancer)

FLUTE project seeks to enhance the diagnosis and treatment of csPCa through the integration of AI and Federated Learning (FL). The project is motivated by the challenges of accessing high-quality medical datasets due to stringent privacy regulations, which limit data sharing across institutions and borders. These challenges lead to biased AI models that are often overfitted to local data and do not generalize well across different populations.

Context and Rationale

In the field of healthcare, particularly in AI-driven medical research, the lack of sufficient and accessible data is a significant limitation. Patient privacy policies, such as the General Data Protection Regulation (GDPR) in Europe, prevent data from being easily shared across borders or institutions. As a result, AI models used for medical diagnosis, especially in PCa, are often trained on localized datasets, which may not capture the full variability of patient characteristics across different regions. This leads to inaccurate models that fail to generalize well when applied to new populations.

FL offers a potential solution by allowing AI models to be trained on data stored at various institutions without moving the data itself. Instead, models are trained locally and then aggregated centrally, ensuring that patient data remains private. However, despite its promise, FL faces challenges related to scalability and privacy, particularly when it comes to sharing data securely across borders.

Hypothesis and Goals

The central hypothesis of the FLUTE project is that integrating AI techniques with FL will lead to the development of powerful predictive tools that can improve the diagnosis of csPCa and predict the aggressiveness of tumors. This approach is expected to facilitate personalized treatment plans that are tailored to the individual characteristics of each patient's tumor, ultimately improving patient outcomes.

The project has three primary objectives:

1. Improved Diagnosis of csPCa: The project aims to enhance the diagnosis of csPCa across Europe by utilizing a robust statistical model. This model, initially developed at the VHIR (Vall d'Hebron Institute of Research) based on seven clinical variables (age, family history of PCa , biopsy type, PSA levels, DRE, prostate volume, and PI-RADS category), will be trained with clinical and imaging data from various regions across Europe, including hospitals in Italy (IRST) and Belgium. By integrating MRI images with clinical data, the model is expected to achieve a higher level of diagnostic accuracy.
2. Development of the FLUTE Platform: The FLUTE platform will be a cross-border federated

AI solution, designed to facilitate secure data sharing and computational analysis among hospitals. By using privacy-enhancing technologies such as Secure Multi-Party Computation (SMPC) and Trusted Execution Environments (TEE), the platform will ensure that sensitive patient data is never exposed during the model training process. This will allow for collaborative research without compromising privacy.

3. Synthetic Data Generation: The project also focuses on generating synthetic data—both clinical and imaging data—to be used for further research. This synthetic data will enable the sharing of research insights without exposing real patient information, thereby facilitating compliance with privacy regulations while still enabling scientific advancement.

Methods and Technologies

The development of the FLUTE platform incorporates a variety of advanced privacy-preserving techniques:

- SMPC allows data to be shared across multiple parties while ensuring that no single party has access to the full dataset.
- TEE provides a secure environment for the execution of computational tasks, ensuring that data remains encrypted even during processing.
- Data Anonymization will be performed on all patient data before it is integrated into the federated system. Each hospital will pseudonymize its data locally, ensuring that individual identities cannot be traced back.

Additionally, secret function sharing and sparse computing will be employed to improve the scalability of the federated algorithms. This will ensure that the system can handle large datasets from multiple hospitals while maintaining computational efficiency. Synthetic data generation will also play a crucial role in ensuring compliance with FAIR data-sharing principles (Findability, Accessibility, Interoperability, and Reusability).

Once the system is developed, the predictive models will be trained and validated using real patient data, and the performance of these models will be compared with those trained on synthetic data. The validation process will involve multiple datasets from the participating hospitals to ensure that the models generalize well across different regions.

Clinical Relevance and Expected Impact

PCa is a common but highly variable disease, with cases ranging from indolent tumors that may not require immediate treatment to aggressive tumors that demand urgent intervention. Currently, there is no universal consensus on how to distinguish between these types, leading to over-diagnosis and over-treatment in many cases.

Recent advancements in medical imaging, particularly mpMRI, have improved the ability to identify csPCa. However, the reliance on traditional markers like PSAD and DRE means that many indolent cases are still being over-diagnosed, while some aggressive cases are missed. The European Association of Urology (EAU) recommends risk-stratified diagnostic pathways, which can reduce unnecessary biopsies and imaging procedures.

The FLUTE project aims to build on these advancements by creating a risk prediction tool that can better distinguish between indolent and aggressive forms of PCa. This will allow for more targeted use of diagnostic procedures, reducing the burden on patients and healthcare systems while improving early detection of aggressive tumors. By using FL, the project will be able to leverage data from multiple European regions, creating a geographically robust model that reflects the diversity of PCa cases across the continent.

Ethical and Legal Considerations

The project is designed to comply with the strict data protection regulations set forth by the GDPR. All data used in the study will be pseudonymized at the local level, with correspondence tables stored separately from the research data to ensure that individual patients cannot be re-identified. In addition, a Data Protection Impact Assessment (DPIA) will be conducted to identify and mitigate any potential risks to patient privacy.

Each participating institution will be responsible for ensuring that data is handled in compliance with local and European regulations. For example, hospitals like VHIR and CHUL will only provide pseudonymized data, and IRST will further anonymize data before it is shared in the federated system. Special attention will be given to the rights of data subjects, with each hospital ensuring that patients are informed of how their data will be used.

Pilot Study and Platform Validation

Before the full implementation of the FLUTE platform, a pilot study will be conducted to assess the feasibility of FL in the context of PCa diagnosis. This pilot study will evaluate the technical infrastructure needed to deploy the platform, test the privacy-preserving methods, and ensure that the system can scale to handle the large datasets expected in the full study.

Once the platform is operational, the AI models will be developed and validated using data from several hospitals. These models will combine clinical data (such as PSA levels, biopsy type, and patient history) with quantitative imaging biomarkers extracted from MRI scans to improve the accuracy of csPCa detection. The models' performance will be assessed based on their accuracy, sensitivity, and specificity, and compared to traditional diagnostic methods.

Outcomes and Contributions

The expected outcomes of the FLUTE project include:

- A functional federated AI platform that enables secure, cross-border data sharing and collaborative model development.
- A validated predictive tool that can be used in clinical settings to assess the probability of csPCa, improving early diagnosis and reducing unnecessary interventions.
- The generation of synthetic datasets that can be used for further research while preserving patient privacy.
- Ethical and legal guidelines for implementing FL in healthcare, ensuring that the platform can be widely adopted in compliance with GDPR.

In summary, the FLUTE project aims to revolutionize PCa diagnosis by combining advanced AI techniques with FL, ensuring patient privacy while enabling the use of large, diverse datasets from across Europe. This approach will lead to more accurate and personalized treatments, improving outcomes for patients with PCa [4].

4.6. IRST contribution to FLUTE project

The FLUTE project represents a significant research endeavor that focuses on advancing the diagnosis and treatment of PCa using cutting-edge technologies such as AI and FL. My involvement, as well as the contribution of the IRCCS IRST, has centered on developing a comprehensive clinical-radiological database and optimizing data management workflows to support the goals of the project.

One of the key challenges we faced at the outset was the manual compilation of clinical-radiological databases for patients undergoing prostate MRI. This process, while necessary for research, was time-consuming and labor-intensive, particularly as it involved gathering data from multiple sources, including patient histories, imaging reports, and pathological findings. With the initiation of the FLUTE project, it became imperative to enhance our database infrastructure to manage large-scale data more efficiently and accurately, especially given the project's focus on integrating information from multiple institutions across Europe.

In collaboration with AGFA HealthCare and the institutional Data Unit, we developed systems to streamline this process. The first step involved automating data extraction from the Radiology Information System (RIS) and integrating it with other clinical sources. This automation allowed us to retrieve essential clinical data, such as PSA levels, biopsy results, and patient histories, directly from the hospital's repository, Log80. The Data Unit played a crucial role in this process, implementing tools to automatically search for patients' MRI images in the PACS and anonymize them for subsequent analysis. This was a critical advancement, as it reduced the manual effort required to locate, extract, and anonymize patient data, ensuring compliance with data protection regulations while preparing the data for use in large-scale research.

My personal involvement in this aspect of the project included providing the initial framework for the database, which contained clinical, anamnestic, radiological, and pathological information for each patient. This initial work laid the groundwork for the development of a dynamic, automated system capable of incorporating new data as it available.

The next phase of the FLUTE project involves the integration of advanced imaging analytics. Starting in late 2024, we will begin analyzing radiological exams using QUIBIM-QP Prostate software to extract quantitative imaging biomarkers from various mpMRI sequences. These biomarkers, which provide detailed insights into tissue characteristics, will be integrated with clinical and laboratory data in a unified database. The goal is to create a robust dataset that combines imaging, clinical, and pathological information to develop predictive models for PCa diagnosis and risk stratification.

By participating in the FLUTE project, IRST has positioned itself at the forefront of research in PCa diagnostics. The integration of FL has the potential to revolutionize the field by enabling the development of highly accurate diagnostic tools while maintaining patient privacy. This aspect of the project is particularly important given the sensitive nature of healthcare data and the stringent regulations governing its use.

The structured databases we are developing will serve as a foundation for these AI-driven models. By harmonizing data from various sources and ensuring that it is anonymized and standardized, we are facilitating the application of AI algorithms that can analyze vast amounts of data to detect patterns and predict patient outcomes with greater accuracy. This approach also enables the IRST to contribute significantly to the collaborative efforts of the FLUTE project, working alongside other leading European institutions to improve PCa diagnosis and treatment on a large scale.

The preliminary results of our work are still in the early stages. So far, we have begun compiling a clinical-radiological database for 400 patients who underwent mpMRI at our institution between 2018 and 2023. This database serves as a starting point for ongoing research and will provide a solid foundation for future developments as we incorporate new data and quantitative imaging biomarkers.

As the FLUTE project moves forward, we anticipate that the predictive models we develop will improve the accuracy of PCa diagnosis, enabling earlier detection and more precise risk stratification. This will ultimately lead to better treatment planning and improved patient outcomes. Furthermore, by leveraging FL and AI technologies, we aim to set new standards in the field of PCa diagnostics, paving the way for more personalized and effective healthcare solutions across Europe.

In conclusion, the FLUTE project offers a unique opportunity for both IRST and myself to contribute to the future of PCa research. Our work in developing automated, standardized databases and integrating advanced imaging analytics is aligned with the project's overarching goals of improving diagnostic accuracy and patient outcomes through innovative AI technologies. As we continue to build on the foundation we have established, we are confident that our contributions will lead to significant scientific advancements and ultimately enhance the quality of care for PCa patients across Europe.

4.7. MRI PRO training course

To harmonize the clinic-radiological multicenter database within the FLUTE project, obtaining expert-level certification in prostate mpMRI reporting was essential. To fulfill this requirement, I successfully completed the MRI PRO course (www.mripro.io), an online Prostate MRI Training Course offered by Monash University. The evidence supporting the use of prostate MRI strongly depends on the experience of specialists who have interpreted a large volume of cases. Given the complexity of prostate MRI, less experienced clinicians may risk missing significant cancers or misinterpreting benign findings, potentially leading to suboptimal patient outcomes. MRI PRO offers clinicians the opportunity to gain substantial experience with immediate feedback, helping them become proficient in prostate MRI interpretation before applying these skills in clinical practice. The primary objective of MRI PRO is to minimize variability and elevate the global standard for accurate prostate MRI reporting. Although I have personally interpreted over 1,000 prostate MRI scans, I recognized the importance of formally testing my competencies and obtaining certification to ensure my expertise aligns with the highest professional standards.

The MRI PRO platform provides access to 300 high-quality, histology-verified prostate MRI cases sourced from international centers. Each case is reviewed through structured reports, and the results are evaluated by a panel of six international expert radiologists. My personal results on 300 cases showed an overall accuracy of 92.8%, which reflects a high level of competence in identifying and reporting PCa lesions.

Completing the MRI PRO course and receiving certification marks a significant milestone in further enhancing my expertise in prostate MRI interpretation. The course allowed me to rigorously validate my skills through the structured analysis of histology-verified cases, accompanied by expert feedback. Achieving an accuracy rate of 92.8% underscores my diagnostic proficiency and reinforces the importance of continuous skill assessment, particularly in a complex and evolving field like prostate MRI. By participating in this comprehensive evaluation, I contributed to the standardization and harmonization efforts for prostate MRI reporting, which are critical for multicenter collaborations like the FLUTE project. Moreover, the certification ensures that I am well-prepared to provide high-quality, consistent results, reducing inter-observer variability and ultimately contributing to improved patient outcomes [5].

5. ML tool for tumor malignancy classification

5.1. Introduction

A collaborative effort between the Medical Physics department and clinical experts was undertaken to bridge specific knowledge gaps and uphold the highest standards of scientific rigor. This interdisciplinary collaboration was essential for addressing the complex technical challenges associated with integrating ML into radiological and clinical workflows. By combining clinical expertise with the specialized knowledge of physics and data science, this partnership facilitated the precise development and rigorous evaluation of the Random Forest (RF) model, ensuring its robust and accurate application in the research context.

The conceptual framework, centered on the clinical and radiological aspects, was developed in conjunction with the Medical Physics team, who played a key role in co-developing the methodology. Their expertise ensured robust data management, algorithmic precision, and rigorous statistical validation, which were crucial for constructing a model capable of delivering reliable clinical insights. The collaboration extended beyond data analysis to include in-depth discussions on dataset selection and result interpretation, ensuring the model's clinical relevance. This interdisciplinary partnership highlights the critical role of integrating clinical knowledge with technical expertise in medical research, leading to more comprehensive and accurate outcomes.

In this chapter, we will detail the dataset used, emphasizing its structure, source, and relevance to the research question. Key preprocessing steps, such as managing missing data, addressing outliers, transforming categorical variables, and normalizing continuous variables, will be outlined. These steps are crucial for ensuring the dataset is in an optimal format for ML applications, preventing distortions that could affect model performance. Additionally, we will discuss the strategies used to balance the dataset, given the typical class imbalance found in medical data, such as the disproportionate representation of benign versus malignant cases. Proper class balancing ensures that the model does not overfit to the majority class, maintaining fairness and accuracy across predictions.

Following the data preparation process, we will describe how the dataset was split into training, validation, and test subsets. This partitioning is critical for evaluating model performance, minimizing overfitting, and ensuring the generalizability of the results to unseen data. By the end of the preprocessing phase, the dataset will be ready for the application of ML algorithms, with the goal of deriving clinically meaningful insights to inform patient management decisions.

The primary focus is the implementation and evaluation of a RF model designed to classify PCa patients based on their clinical and radiological data. Specifically, the model aims to predict the ISUP grades. Stratifying patients by their ISUP grade is crucial for guiding treatment decisions, such as any treatment for healthy patients (ISUP 0), opting for active surveillance in less aggressive cases (ISUP 1) or more intensive interventions like surgery or radiotherapy for more severe cases (ISUP 2+). The objective of this analysis is to accurately distinguish between ncsPCa (ISUP 0 and 1) and csPCa (ISUP 2+).

ML models, such as RF, have gained prominence in medical fields, especially in supporting complex decision-making tasks. Radiologists and clinicians often face the challenge of integrating vast amounts of data, ranging from mpMRI to clinical biomarkers such as PSA levels and patient history. ML models provide a systematic, data-driven approach to handle these complexities, enabling the identification of patterns that may not be easily detectable by human experts. In particular, RF models, which leverage ensemble learning, offer robustness in handling heterogeneous data sources, making them ideal for applications in PCa diagnosis.

One of the key advantages of ML in radiology is its ability to quantify subtle imaging features, such as lesion shape, texture, and intensity, which might elude even experienced clinicians. Radiomics, a growing field in medical imaging, seeks to extract these quantitative features from radiological images, providing new avenues for cancer diagnosis and treatment planning. As demonstrated by Gillies et al. (2016), radiomics can enhance the prediction of disease outcomes, aiding in the development of personalized treatment strategies. In PCa, mpMRI is a rich source of information, and integrating ML models with radiomics has the potential to improve diagnostic accuracy, supplementing clinical expertise with objective, reproducible predictions.

However, while ML models hold great promise, their integration into clinical practice must be approached cautiously. These models are not intended to replace clinical judgment but to support it. Physicians must critically evaluate the predictions generated by these models, considering their limitations, such as biases in training data or issues with generalizability to different patient populations. Moreover, ML models, while proficient at detecting patterns, lack the contextual awareness that comes from years of clinical experience. This makes the physician's role indispensable in validating and interpreting the model’s outputs, ensuring they align with the broader clinical picture.

In conclusion, while ML models like RF offer significant advantages in enhancing diagnostic accuracy and improving patient care, their use in clinical practice should be seen as a collaborative tool, augmenting but not replacing the expertise of clinicians. As these models continue to evolve and prove their utility in areas like PCa classification, their role in guiding clinical decisions will likely expand, provided they are used responsibly and in conjunction with sound medical judgment [6-7-8-9].

5.2. Dataset Features

The dataset includes a combination of clinical and radiological features, essential for classifying PCa severity. Below is an updated and detailed description of the key independent variables, considering the specific methodologies used for the PI-RADS and ISUP scores, as well as the method for calculating the rADC feature.

- PSA: PSA is a blood biomarker commonly used in PCa screening. In this dataset, the PSA value corresponds to the blood sample taken closest to the date of the MRI scan, ensuring it reflects the most current information about the patient's PCa status at the time of imaging. Elevated PSA levels are associated with an increased likelihood of PCa, although they can also be elevated in benign conditions such as BPH or prostatitis. PSA levels are crucial in assessing PCa risk, particularly when considered alongside other clinical and radiological features.

- Prostate Volume: The prostate volume is calculated using three diameters—anteroposterior (AP), laterolateral (LL), and craniocaudal (CC)—measured on the T2w MRI sequence. These dimensions are multiplied by 0.52, following the volumetric model of a sphere, to estimate the prostate volume. However, advanced imaging software with automatic prostate contouring capabilities offers a more precise and accurate assessment of prostate volume. Once these technologies become available in our clinical practice, we intend to incorporate them for enhanced volumetric accuracy. This method provides an approximation of the gland's size and is critical for evaluating prostate health. Larger prostate volumes can naturally elevate PSA levels, making this feature essential for calculating PSAD.

- PSAD: PSA density is calculated by dividing the PSA level by the prostate volume. It provides a more accurate assessment of PCa risk than PSA alone, particularly in patients with larger prostate volumes, where PSA levels may be elevated due to benign factors.

- Age: The age of the patient at the time of the MRI exam. Age is a well-known risk factor for PCa, with increasing age correlating with a higher likelihood of developing the disease.

- PI-RADS Score: The PI-RADS score is used to evaluate prostate lesions based on mpMRI. In this dataset, the PI-RADS score corresponds to the lesion with the highest score, as determined by the imaging. If multiple lesions have the same PI-RADS score, the lesion with the largest volume is selected as the dominant lesion. PI-RADS scores range from 1 to 5, with higher scores indicating a higher likelihood of csPCa.

- rADC (pathological/healthy): This feature represents the ratio of ADC values between pathological and healthy tissues. The ADC values are derived from DWI, which measures the movement of water molecules in tissues. For the rADC calculation, two regions of interest (ROIs) are defined: one in the cancerous region and the other in the contralateral healthy tissue at the same level and in the same zone (e.g., CZ with CZ, PZ with PZ). The rADC is the ratio of the ADC values between these two ROIs, providing a quantitative measure to differentiate cancerous tissue from healthy tissue. This study was conducted retrospectively, utilizing previously acquired imaging data. Lower ADC values in the pathological ROI compared to the healthy ROI typically indicate higher cellular density and more aggressive tumors.

- Zone**:** This feature indicates the anatomical zone of the prostate in which the dominant lesion is located:
  - **N** = for negative patients with no lesions.
  - **C** = for lesions located in the CZ, TZ, or anterior portion (AS).
  - **P** = for lesions located in the PZ.

5.3. Target Variable (ISUP)

ISUP Class (0, 1, 2+): The target variable in this study is based on the highest GS found in either histological examination or biopsy results. If histological data from surgery or resection is available, the ISUP class reflects the highest GS provided by the pathologist. If histology is not available (e.g. patient treated with radiotherapy) , the highest GS from the biopsy sample is used. The ISUP classification system is a key determinant of cancer aggressiveness:

- ISUP 0: No PCa.
- ISUP 1: ncsPCa, typically managed through active surveillance.
- ISUP 2+: csPCa requiring immediate therapeutic intervention, such as surgery or radiotherapy.

This comprehensive dataset, which combines clinical variables (PSA, prostate volume, age) with imaging features (PI-RADS score, rADC), allows for a robust assessment of PCa risk and severity, aiding in diagnostic accuracy and treatment decision-making.

5.4. Data Preprocessing

Data Cleaning:

Data cleaning is a critical step in any ML pipeline, particularly in medical datasets where missing values and outliers can significantly affect the model's performance and the validity of predictions.

- *Handling Missing Values*: Missing data was addressed systematically to ensure no important information was lost, while still maintaining the integrity of the dataset. For numerical variables such as PSA, prostate volume, and rADC, statistical imputation techniques were used, often replacing missing values with the median to avoid the influence of extreme values (which can affect the mean). This approach ensures that each data point is as complete as possible, reducing the chance of bias in the model.

- *Outlier Detection and Management*: Medical datasets often contain outliers that can either represent true rare cases or erroneous data points. In this study, outliers were identified through statistical methods such as Z-scores and interquartile ranges (IQR). It was important to differentiate between biological variations (legitimate outliers) and data entry errors. For instance, extremely high PSA levels might indicate aggressive PCa but can also result from technical issues or inaccuracies in measurement. Outliers that were determined to be biologically plausible were retained, while those likely due to errors were either corrected or removed.

Feature Selection:

Feature selection plays a pivotal role in reducing the complexity of the model, minimizing overfitting, and improving interpretability (particularly crucial in clinical settings where the decision-making process must be transparent and justifiable) .

- *Motivation for Feature Selection*: The primary features included in the model were selected based on their clinical relevance to PCa. Key features such as PSA, prostate volume, rADC, PI-RADS score, and Zone have established roles in the literature for predicting PCa severity. Additionally, PSAD was derived by dividing PSA by prostate volume, as this metric provides a more refined assessment of PCa risk, particularly in cases where PSA alone may be misleading due to large prostate volumes.

- *Avoiding Multicollinearity*: While feature selection is guided by clinical insight, multicollinearity between variables can degrade model performance by inflating variance. Therefore, the correlation matrix of the selected features was analyzed to ensure that highly correlated features did not coexist in the model, which could lead to redundancy. For example, prostate volume and PSAD are related, but PSAD was preferred as it normalizes PSA relative to prostate size, improving its predictive value.

- *Dimensionality Reduction*: Although this dataset contained a manageable number of features, it is important to consider dimensionality reduction techniques, such as principal component analysis (PCA), when dealing with higher-dimensional datasets. These methods can reduce feature complexity while retaining most of the variance, though in this case, the clinically meaningful features were sufficient for building an interpretable and effective model.

Dataset Split:

Proper division of the dataset into training and test sets is critical to evaluate model performance and ensure generalizability. In this study, the dataset was split into a training set (70%) and a test set (30%).

- *Training and Test Split*: The choice to use 70% of the data for training allows the model to learn from a sufficiently large portion of the data while reserving 30% for testing ensures that there is a robust evaluation of the model's performance. This test set is used to validate the model's ability to generalize to unseen data, a critical factor in clinical applications where predictions must be accurate across a broad patient population.

- *Selection of random_state*: The random seed, or random_state, was fixed to ensure reproducibility of the results. Reproducibility is essential in scientific studies, particularly in medicine, where decision-making relies on replicable outcomes. A fixed random_state allows researchers to verify that the model behaves consistently across different runs and environments.

- *Test Size*: A 30% test size provides a good balance between having enough data to train the model effectively and retaining a sufficient portion of data to thoroughly evaluate the model's performance. In smaller medical datasets, this split might be adjusted to ensure that rare cases (e.g., aggressive cancer cases) are adequately represented in both the training and test sets. However, in this study, 30% was sufficient to capture the distribution of ISUP grades across the patient population.

5.5. Feature Preprocessing

Standardization of Numerical Variables:

Even though RF is not affected by the scale of features (since it uses decision trees that rely on feature splits rather than distance-based calculations), there are still scenarios where standardization can be useful for consistency and interpretability, especially when working in conjunction with other models or when analyzing feature importance.

- *Importance of Standardization*: Numerical variables such as PSA, rADC, prostate volume, and PSAD have varying scales and units of measurement. While this variation does not directly affect the RF algorithm, standardization can be useful for visualizing feature importance on a comparable scale and improving the consistency of the preprocessing pipeline, especially if the same dataset might later be used for other ML algorithms.

- *Application of StandardScaler*: Standardization was applied using the StandardScaler from the sklearn.preprocessing library, which centers the data by subtracting the mean and scales it by the standard deviation. This ensures that all numerical variables have a mean of 0 and a standard deviation of 1. The following features were considered for standardization:
    - PSA
    - rADC
    - Prostate Volume (measured in ml)
    - PSAD (measured in ng/ml/cm^3)

Although this step isn't essential for RF itself, it adds uniformity across the pipeline when comparing feature importances or experimenting with different algorithms.

Handling of Categorical Variables:

In the dataset, categorical variables like PI-RADS score and Zone play a key role in the classification process.

- *Encoding Techniques*: Since RF can handle both numerical and categorical data (through splitting decisions based on thresholds for numerical features and categorical values for categorical features), categorical encoding was necessary to convert non-numeric categorical variables into a format usable by the model. The PI-RADS and Zone variables were handled as follows:
    - PI-RADS: This ordinal variable, which ranges from 1 to 5, was label encoded, as the values have an inherent order where higher PI-RADS scores indicate a higher likelihood of csPCA.
    - Zone: This categorical variable represents the location of the lesion in the prostate, such as N (no lesions), C (CZ,TZ,AS), and P (PZ). To preserve the distinct categories without imposing any ordinal relationship, One-Hot Encoding was applied to convert each zone into binary indicator variables. For instance, each patient would have separate columns indicating whether their lesion is located in the peripheral or central regions. ('C' = 0, 'N' = 1, 'P' = 2)

The choice of One-Hot Encoding ensures that the categorical variable does not imply a false order or ranking, which could mislead the model.

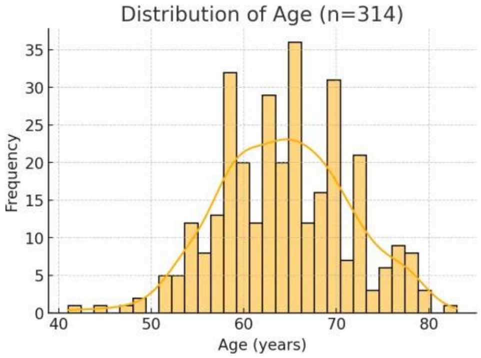
By properly encoding categorical variables and handling numerical features, the dataset was made suitable for use in the RF classifier, ensuring that both feature types were appropriately considered during training.

5.6. Data presentation

This chapter provides a comprehensive overview of the cleaned dataset used to train the ML model for tumor malignancy classification. The dataset includes several key demographic, clinical, and radiological variables essential for building an accurate predictive tool.

Demographic Variables

- **Age**: The histogram displays the age distribution of 314 patients, with a majority of individuals aged between 55 and 75 years. The data shows a normal distribution centered around the 60–65 year range, with fewer patients at the extremes of the age spectrum. The orange line represents the smoothed density curve, illustrating the underlying distribution of age among the cohort.



Age distribution of the patients in the dataset

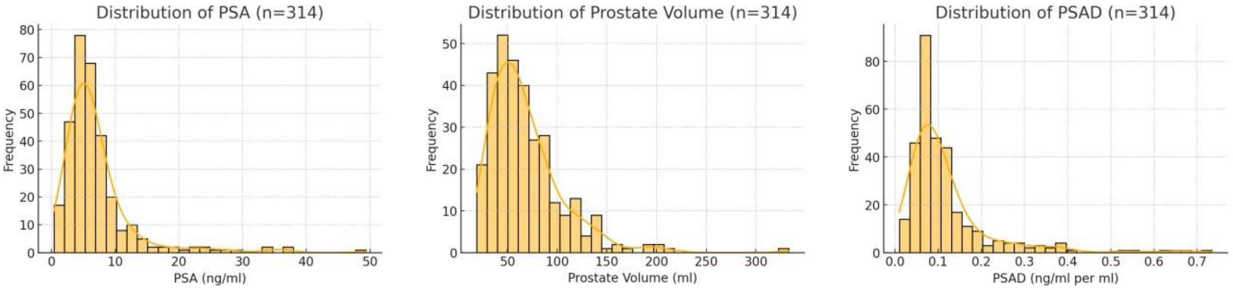
Clinical Data

Clinical variables, including PSA, prostate volume, and PSAD, are distributed in a skewed log-normal distribution.

- **PSA**: Distribution of PSA (ng/ml) levels shows a right-skewed distribution with the majority of values below 10 ng/ml.

- **Volume**: Distribution of Prostate Volume (ml) reveals that most patients have prostate volumes ranging between 30 and 100 ml.

- **PSAD**: Distribution of PSAD (ng/ml/cm³) shows a skewed distribution, with most values concentrated below 0.2, indicating a higher concentration of PSA relative to prostate volume in some patients.

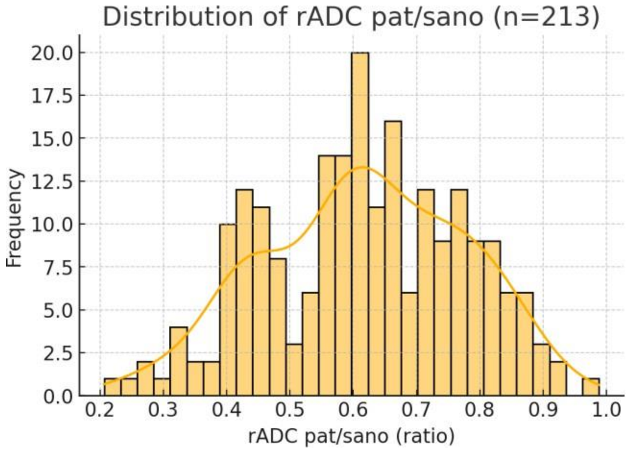


Distribution of PSA, Prostate Volume, and PSAD.

Radiological Data

Radiological features calculated from multiparametric MRI exams are crucial for the model's predictive capabilities.

- **rADC**: The histogram shows the distribution of rADC values for 213 positive patients. The rADC ratio is calculated by dividing the ADC value of pathological tissue by that of healthy tissue. The majority of rADC ratios fall between 0.4 and 0.7, indicating the relative diffusion of water molecules in pathological versus healthy tissue. The orange line represents the smoothed density curve, highlighting the underlying distribution of the data.

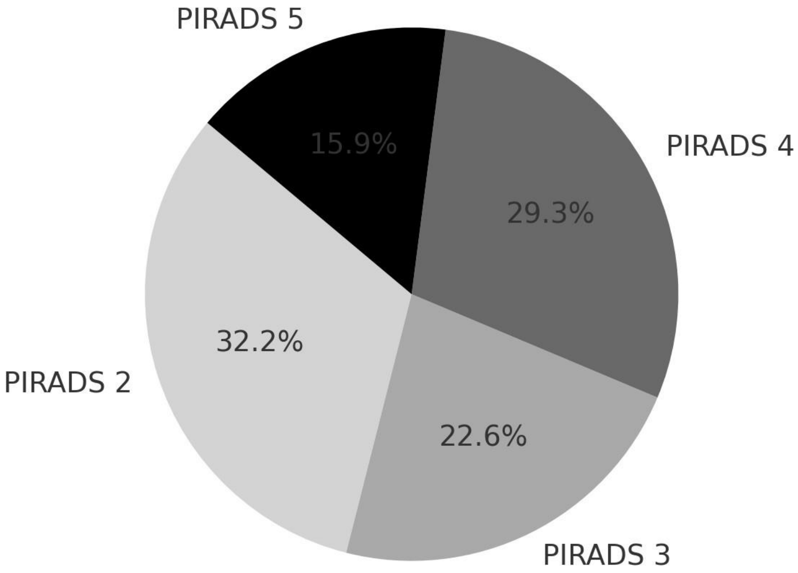


Distribution of the rADC (pathological/healthy tissue) ratio.

- **PI-RADS Score**: The distribution of patients across PI-RADS categories in this cohort reveals that PI-RADS 2 has the highest representation with 101 patients (32.2%), followed by PI-RADS 4 with 92 patients (29.3%). PI-RADS 3 includes 71 patients (22.6%), while PI-RADS 5, representing more aggressive lesions, comprises 50 patients (15.9%).

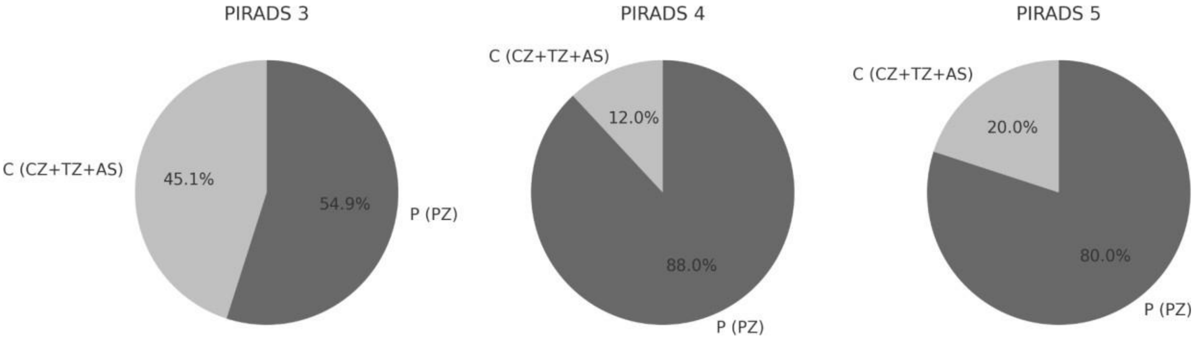
	PATIENTS	PERCENTAGE
PI-RADS 2	101	32,2%
PI-RADS 3	71	22,6%
PI-RADS 4	92	29,3%
PI-RADS 5	50	15,9%

PI-RADS Patient Distribution by Category



Proportional Distribution of Patients by PI-RADS Category

- **Zone:** The pie charts show the proportion of lesions located in the PZ versus the other (CZ + TZ + AS) for each PI-RADS score. Higher PI-RADS scores are predominantly associated with lesions in PZ, with a marked increase in proportion as the PI-RADS score rises from 3 to 5.



Distribution of PI-RADS Scores 3, 4, and 5 across different prostate zones

This table summarizes the distribution of lesion localization according to the PI-RADS score classification. The columns show the different lesion zones: N (No lesion), C (CZ + TZ + AS), and P (PZ). PI-RADS 2 primarily consists of patients with no detected lesions, while PI-RADS 3, 4, and 5 reveal a progression of lesion occurrence from the central to peripheral zones, with PI-RADS 5 showing the highest number of lesions in the PZ.

	LESION LOCALIZATION		
PI-RADS SCORE	**N (No lesion)**	**C (CZ+TZ+AS)**	**P (PZ)**
P2	101	0	0
P3	0	32	39
P4	0	11	81
P5	0	10	40

Lesion Localization Based on PI-RADS Score.

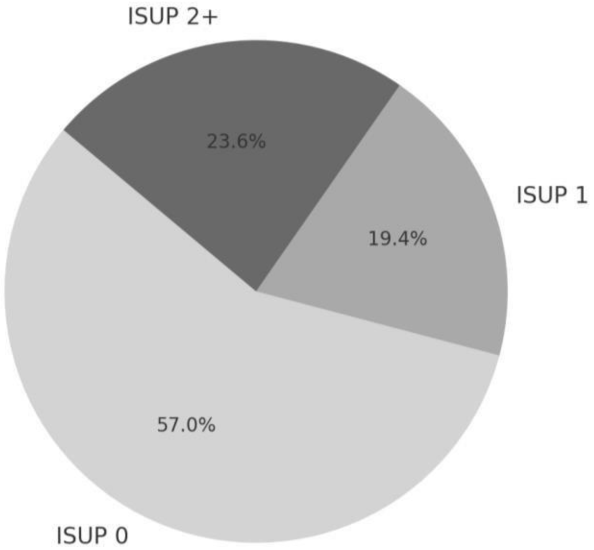
Histopathological Data

The histopathological data analyzed here derive from prostatectomies and prostate biopsies, providing insights into prostate tissue characteristics and lesion grading. This data supports diagnostic accuracy and informs treatment strategies for prostate-related conditions.

- **ISUP Classification:** The histopathological data reveal a distribution across ISUP grades, with 179 patients classified as ISUP 0, 61 patients as ISUP 1, and 74 patients as ISUP 2 or higher.

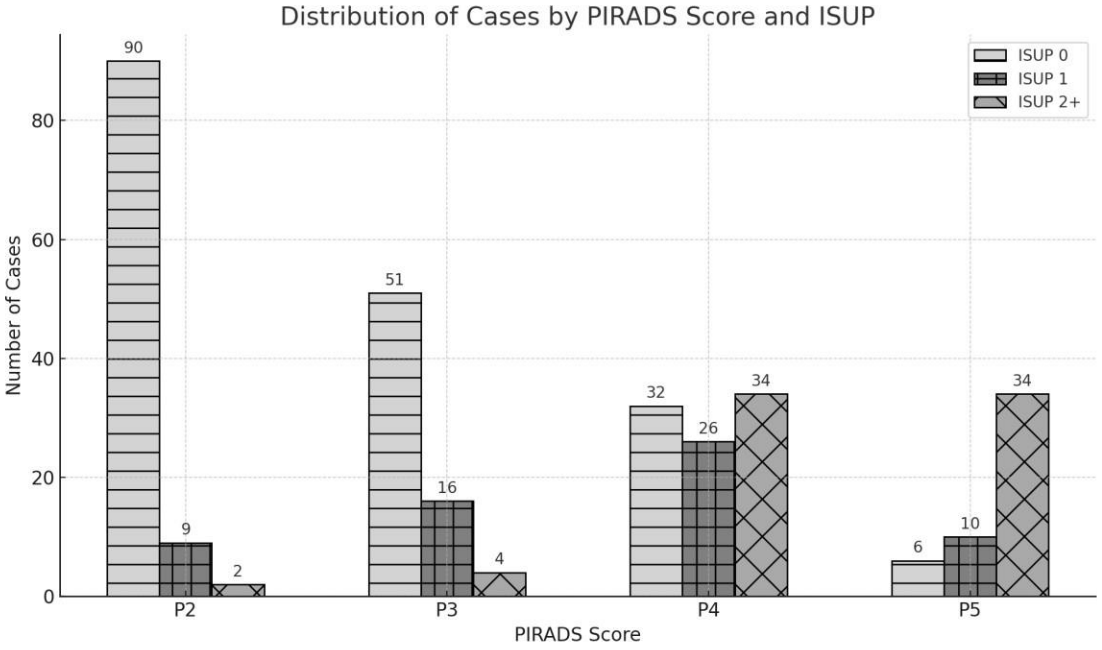
	PATIENTS	**PERCENTAGE**
ISUP 0	179	57%
ISUP 1	61	19,4%
ISUP 2+	74	23,6%

ISUP Patient Data



Proportional Distribution of Patients by ISUP Category

- **ISUP Classification and PI-RADS Score :** This bar chart shows the number of cases stratified by PI-RADS score and ISUP classification (ISUP 0, ISUP 1, ISUP 2+). The majority of PI-RADS 2 cases fall under ISUP 0, indicating no malignancy, while higher PI-RADS scores (P4 and P5) tend to have a greater distribution in ISUP 2+, representing higher malignancy risk.



Distribution of Cases by PI-RADS Score and ISUP Classification

The table summarizes the distribution of ISUP grades (0, 1, 2+) across different PI-RADS scores (P2, P3, P4, P5). Each cell represents the number of cases within a specific PI-RADS score and ISUP grade category. The totals in the rightmost column and bottom row indicate the sum of cases for each ISUP grade and PI-RADS score, respectively.

	PI-RADS SCORE				**TOTAL**
	P2	**P3**	**P4**	**P5**	
ISUP 0	90	51	32	6	**179**
ISUP 1	9	16	26	10	**61**
ISUP 2+	2	4	34	34	**74**
TOTAL	**101**	**71**	**92**	**50**	**314**

Distribution of ISUP Grades Across PI-RADS Scores

Correlation Matrix analysis

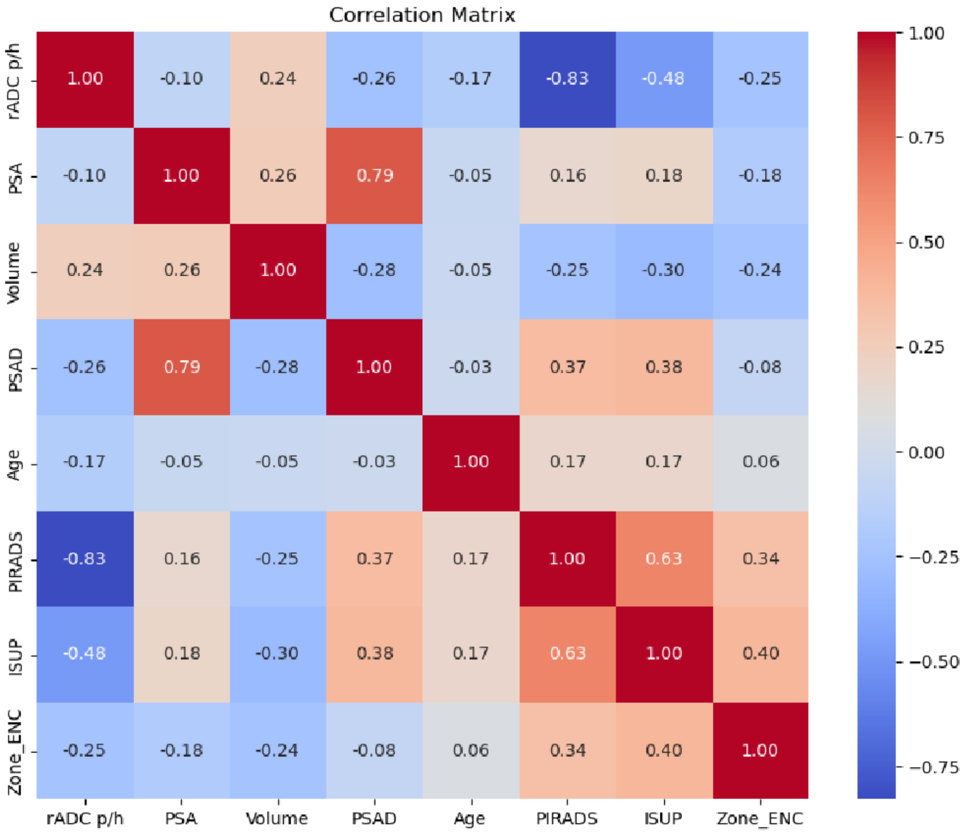
The correlation matrix provides insight into the relationships between various clinical, radiological, and pathological variables, specifically focusing on the ISUP grade, which indicates the aggressiveness of PCa. Understanding these correlations aids in identifying predictive factors for cancer severity and guides clinical decision-making.

- ISUP and PI-RADS (0.63): The strongest positive correlation observed is between ISUP grade and PI-RADS score (0.63). This significant relationship implies that higher PI-RADS scores, which reflect higher suspicion of malignancy on MRI, are often associated with higher ISUP grades determined through histopathology. Clinically, this correlation underscores the value of PI-RADS in predicting tumor aggressiveness and supports its use as a non-invasive imaging tool in assessing cancer severity.

- ISUP and rADC (-0.48): ISUP grade has a moderate negative correlation with the rADC value (-0.48). This inverse relationship suggests that lower ADC values, which indicate restricted diffusion on imaging, are associated with higher ISUP grades. Restricted diffusion is a common characteristic of high-grade tumors, making rADC an essential biomarker in identifying aggressive cancer. The strong negative correlation reinforces the utility of DWI in evaluating PCa risk.

- ISUP and PSAD (0.38): The ISUP grade shows a moderate positive correlation with PSAD at 0.38, suggesting that PSAD may be associated with higher ISUP grades. This correlation is stronger than that of total PSA, which has a much lower correlation of 0.18 with ISUP, indicating that PSAD is a more reliable indicator of cancer aggressiveness than PSA alone. Given PSA's role as a widely-used biomarker, these findings emphasize the added value of PSAD in assessing overall cancer risk, especially when used alongside imaging and histopathological data.

In summary, the correlations observed in these data provide valuable insights for our model aimed at predicting PCa aggressiveness. The strong positive relationship between ISUP and PI-RADS (0.63) demonstrates the predictive power of PI-RADS in non-invasively assessing tumor severity, making it a key feature for our model. The moderate negative correlation between ISUP and rADC (-0.48) underscores the importance of DWI parameters, such as rADC, as indicators of high-grade tumors. Additionally, the positive correlation between ISUP and PSAD (0.38) underscores their value as complementary features. Together, these variables form a comprehensive dataset that enhances our model's ability to accurately predict cancer aggressiveness, guiding personalized diagnostic and treatment decisions.


Correlation Matrix

	rADC p/h	PSA	Volume	PSAD	Age	PIRADS	ISUP	Zone_ENC
rADC p/h	1.00	-0.10	0.24	-0.26	-0.17	-0.83	-0.48	-0.25
PSA	-0.10	1.00	0.26	0.79	-0.05	0.16	0.18	-0.18
Volume	0.24	0.26	1.00	-0.28	-0.05	-0.25	-0.30	-0.24
PSAD	-0.26	0.79	-0.28	1.00	-0.03	0.37	0.38	-0.08
Age	-0.17	-0.05	-0.05	-0.03	1.00	0.17	0.17	0.06
PIRADS	-0.83	0.16	-0.25	0.37	0.17	1.00	0.63	0.34
ISUP	-0.48	0.18	-0.30	0.38	0.17	0.63	1.00	0.40
Zone_ENC	-0.25	-0.18	-0.24	-0.08	0.06	0.34	0.40	1.00

5.7. Model Selection

The dataset used for classifying PCa patients based on ISUP grades (0, 1, 2+) contains a combination of clinical and radiological variables. The main features include rADC, PSA, prostate volume, PSAD, calculated age, PI-RADS, and ZONE. These variables present several challenges due to their heterogeneous nature (both continuous and categorical data types) and the presence of non-linear relationships. Below is a detailed explanation of why RF is the most suitable model for this dataset, compared to other ML algorithms.

1. **Random Forest (RF)**

Pros:

- Handling heterogeneous data: RF is highly effective with datasets that combine both numerical and categorical variables, like ours. It requires minimal preprocessing and can easily handle continuous variables like PSA and PSAD, as well as categorical variables such as PI-RADS and Zone, without the need for extensive transformations such as one-hot encoding.

- Reducing overfitting: RF operates as a bagging algorithm that builds multiple decision trees using different data subsets, reducing the risk of overfitting. This is particularly advantageous in our dataset, where we have a relatively limited sample size (314 patients) with a large number of features, reducing the risk of overfitting due to the model's ability to generalize.

- Feature importance: RF provides insight into the importance of each feature in the classification task. For example, it allows us to evaluate whether rADC, PSA, or PI-RADS has the most significant impact on predicting ISUP grades. This is crucial for clinical applications, where understanding the most influential features can guide better decision-making.

- Handling class imbalance: In the dataset, cases of high-risk PCa (ISUP 2+) are less frequent compared to lower-risk cases (ISUP 0 and 1). RF can assign weights to classes to address this imbalance, ensuring that aggressive cancer cases are not underrepresented in the model's predictions.

Cons:

- Computational costs: Training a RF model with many trees and high-dimensional features (as in our case with several radiological and clinical variables) can be computationally expensive. However, with sufficient computational resources and parallel processing, this issue can be mitigated.

Conclusion:

RF is highly suitable for our dataset due to its ability to handle heterogeneous data, reduce overfitting, and provide valuable insights into feature importance. Its flexibility in dealing with imbalanced classes and complex feature interactions makes it an ideal choice for classifying PCa based on ISUP grades.

2. **Logistic Regression (LR)**

Pros:

- High interpretability: LR provides interpretable coefficients that show the impact of each feature on the prediction outcome. This is valuable in our context, where we can clearly see the influence of variables such as PSA or PI-RADS on the likelihood of a patient having aggressive PCa (ISUP 2+).

Cons:

- Limited to linear relationships: LR assumes a linear relationship between independent variables and the predicted class. In our dataset, however, there are non-linear interactions between clinical and radiological variables. For example, the combined effect of PSA and PI-RADS on PCa progression cannot be effectively modeled with a linear relationship.

- Poor performance with high-dimensional data: Our dataset includes multiple complex radiological features, and LR may not perform well without significant feature engineering or dimensionality reduction, which could result in lost information.

Conclusion:

While LR offers simplicity and interpretability, its limitations in modeling non-linear relationships and handling complex, high-dimensional data make it less suitable for our task.

3. **Support Vector Machine (SVM)**

Pros:

- Effective in high-dimensional spaces: SVM performs well in datasets with many features and can effectively separate classes using non-linear kernel functions. This is advantageous for our dataset, which includes complex imaging-derived features from mpMRI scans, such as rADC and PI-RADS.

Cons:

- Sensitive to class imbalance: SVM tends to underperform when classes are imbalanced, as in our dataset, where ISUP 2+ cases are underrepresented. This may result in poor performance in identifying aggressive PCa cases unless special techniques are used, such as adjusting class weights.

- High computational cost: Training SVMs, especially with non-linear kernels such as RBF, can be computationally expensive and time-consuming. Given our dataset with high-dimensional radiological data, this could become a practical limitation.

Conclusion:

SVM could handle the complexity of radiological data well, but its sensitivity to class imbalance and computational expense make it less practical for this specific task.

4. **K-Nearest Neighbors (KNN)**

Pros:

- Non-parametric: KNN makes no assumptions about the underlying data distribution, which can be beneficial for heterogeneous datasets like ours.

Cons:

- Curse of dimensionality: KNN suffers from the curse of dimensionality. In datasets with many features like ours, the distances between data points become less meaningful, leading to poor classification performance. For example, subtle differences in rADC or PI-RADS might not be accurately captured.

- Difficulty with class imbalance: KNN relies on the majority of neighbors for classification, meaning that if the majority class dominates (e.g., ISUP 0), it is likely to bias the predictions toward this class, underrepresenting the more aggressive cancer cases (ISUP 2+).

Conclusion:

KNN is unsuitable for our dataset due to the high number of features and the imbalanced class distribution, which are likely to reduce its classification accuracy.

5. **Neural Networks (NN)**

Pros:

- Ability to model complex, non-linear relationships: NN excel at capturing complex interactions between variables. This would be advantageous for our dataset, where there are intricate relationships between features such as rADC, PI-RADS, and PSA, which other models may struggle to capture.

Cons:

- Requires large datasets: NN perform best when trained on large amounts of data. With only 314 samples, our dataset may be too small for a neural network to generalize effectively, leading to overfitting.

- Lack of interpretability: NN are often described as "black-box" models. In clinical contexts, it is critical to understand and explain why a certain prediction was made, which is difficult with NN.

Conclusion:

While NN are powerful in modeling complex relationships, the small size of our dataset and the need for interpretability in a clinical setting make them less practical for this task.

6. **Conclusion**

Given the structure and challenges of our PCa dataset, RF stands out as the most suitable model. Its ability to handle both clinical and radiological data types, prevent overfitting in moderately-sized datasets, manage class imbalance, and provide interpretable feature importance rankings makes it the optimal choice for predicting ISUP grades based on clinical and mpMRI-derived features.

5.8. RF Model

Description of the Algorithm:

RF is an ensemble learning algorithm that combines multiple decision trees to perform classification or regression tasks. It is based on the idea of aggregating the predictions of many individual decision trees to create a more accurate and robust model. In classification problems, each decision tree in the RF votes for a class, and the final prediction is determined by majority voting among all the trees.

The algorithm works by constructing a multitude of decision trees, each trained on a random subset of the training data, a technique known as bagging (bootstrap aggregating). Additionally, during the construction of each tree, a random subset of features is chosen for each split, adding more randomness and helping to decorrelate the trees. This randomness makes RF less prone to overfitting compared to individual decision trees, which tend to overfit the training data.

RF is particularly effective for multiclass classification problems because it can handle non-linear relationships, mixed data types (numerical and categorical), and automatically assess feature importance. This makes it a versatile and powerful tool in medical data analysis, where the relationships between features can be complex and nonlinear.

Advantages of Using RF for Multiclass Classification:

- Robustness to Overfitting: By averaging the predictions of multiple decision trees, RF reduces overfitting. While a single decision tree might overfit the training data, RF mitigates this by introducing randomness in data sampling and feature selection, making the model more generalizable.

- Handling Non-Linear Relationships: Unlike linear models, RF is capable of capturing non-linear relationships between features, which is especially useful for medical data, where interactions between variables such as PSA, PI-RADS, and prostate volume are often non-linear.

- Effective for Mixed Data Types: RF can seamlessly handle both numerical and categorical data without extensive preprocessing. This makes it suitable for medical datasets, where different types of data (e.g., imaging scores, clinical measurements, and categorical labels) are commonly found.

- Feature Importance: RF provides a measure of feature importance based on how useful each feature is in improving splits across all trees. This feature is invaluable in medical research, as it provides insights into which variables are most important in predicting outcomes, such as distinguishing between different PCa grades.

- Multiclass Capabilities: For the multiclass problem of classifying ISUP grades (0, 1, 2+), RF can naturally handle multiple classes without requiring complex modifications to the algorithm. This is particularly advantageous in comparison to other algorithms that require special extensions for multiclass classification.

Initial Model Configuration:

The RF model used in this analysis was configured with default parameters, with a few important adjustments to improve performance, particularly given the challenges presented by class imbalance:

- Number of Trees (n_estimators): The default value of 100 trees was used to start the modeling process. This number of trees is generally sufficient to capture complex relationships in the data while maintaining computational efficiency. Increasing the number of trees might slightly improve accuracy, but with diminishing returns in terms of computational cost.

- Max Features (max_features): By default, RF selects the square root of the number of features for each split. This ensures that individual trees are diverse, as they use different subsets of features, which helps in reducing correlation among trees and thereby enhances the model's robustness.

- Class Weight (class_weight='balanced'):
  - In this dataset, the target variable, ISUP class, is not evenly distributed across the classes. The majority of patients fall into ISUP grades 0 and 1, while ISUP 2+ (which indicates more aggressive cancer) is underrepresented. This class imbalance poses a challenge, as the model might become biased towards predicting the more frequent classes, neglecting the minority class, which is often the most clinically significant.
  - The class_weight='balanced' parameter was used to address this issue. This option automatically adjusts the weights of the classes inversely proportional to their frequencies in the training dataset. As a result, the model gives more emphasis to minority classes (such as ISUP 2+), making sure that they are well-represented in the training process. This adjustment helps to improve the model's sensitivity to the minority class, ensuring that aggressive cancer cases are accurately identified.

- Bootstrap (bootstrap=True): By default, RF uses bootstrap sampling to create different training subsets for each tree. This approach ensures that each tree is built on a slightly different subset of the data, which contributes to reducing overfitting and increasing the generalizability of the model.

Conclusion:

The RF algorithm is particularly well-suited for medical data classification problems, such as predicting PCa severity (ISUP grades), due to its ability to handle complex, non-linear interactions among features, robustness against overfitting, and capacity to deal with mixed data types. The use of default parameters, such as 100 trees and square root of features for splitting, provided a good starting point, while class_weight='balanced' played a crucial role in addressing the inherent class imbalance, ensuring the model's sensitivity to less frequent but clinically significant cases. These characteristics make RF an effective choice for predicting outcomes in clinical datasets, where accuracy and reliability are paramount.

Hyperparameter Optimization

Motivation:

Hyperparameter optimization is a crucial step in improving the performance of ML models. Unlike model parameters, which are learned from the data during training, hyperparameters are external settings that guide how the model is trained. For example, the number of trees in a RF or the depth of each tree are hyperparameters that significantly impact the model's performance.

Optimizing these hyperparameters can enhance the model's ability to generalize to unseen data, thereby avoiding both underfitting and overfitting. In medical contexts, such as predicting PCa severity using ISUP grades, hyperparameter tuning is particularly important because it helps improve the accuracy, precision, and robustness of the predictions, which can directly influence clinical decision-making.

Techniques Used:

GridSearchCV:

- To perform hyperparameter optimization, GridSearchCV from the sklearn.model_selection library was used. This technique systematically evaluates all combinations of hyperparameter values defined in a predefined grid. The purpose is to find the combination that yields the best model performance. GridSearchCV is exhaustive, making it suitable for smaller hyperparameter grids where a thorough search is possible.

- Parameters Considered: In this study, several important hyperparameters for the RF model were considered:
  - Number of Estimators (n_estimators): This parameter controls the number of decision trees in the forest. Values such as [100, 200, 300] were tested to understand the impact of adding more trees on model performance.
  - Max Depth (max_depth): This limits how deep each tree in the forest can grow. Values such as [10, 20, 30, None] were explored to control model complexity. Limiting depth prevents overfitting by making trees simpler.
  - Minimum Samples per Split (min_samples_split): Defines the minimum number of samples required to split an internal node. Values like [2, 5, 10] were tested to control how the tree grows, which can significantly affect overfitting and underfitting.
  - Minimum Samples per Leaf (min_samples_leaf): This parameter represents the minimum number of samples required to be at a leaf node, with values like [1, 2, 4] considered. Increasing this parameter can help in creating more generalized and stable models.
  - Maximum Features (max_features): Controls the number of features considered when looking for the best split. Values such as ['sqrt', 'log2'] were included to find the optimal number of features, balancing tree diversity with depth.

- RandomizedSearchCV (Optional): When the hyperparameter search space is large, RandomizedSearchCV is often used as an alternative to GridSearchCV. It selects random combinations of parameters from the defined ranges, which can be computationally more efficient for a large number of combinations while still providing good results.

Cross-Validation:

- Cross-Validation Process: To evaluate model performance during hyperparameter tuning, K-Fold Cross-Validation was used with k=5. This means that the training data was split into five equal parts. The model was trained on four parts and tested on the remaining part, repeating this process five times so that each part was used for testing once. The final model performance was then averaged over these five iterations. This approach helps in estimating the model's ability to generalize to unseen data and provides a more reliable assessment of model performance compared to a single train-test split.

- Stratified Cross-Validation for Imbalanced Classes: Given that the ISUP classes are not evenly distributed in the dataset, a stratified version of K-Fold Cross-Validation was used. In stratified cross-validation, each fold maintains the same class distribution as the original dataset. This is particularly important in medical datasets where one or more classes may be underrepresented. By using stratified cross-validation, the model is exposed to a similar distribution of classes in each fold, ensuring that minority classes are adequately represented during training and evaluation. This is essential for avoiding biased models that over-predict the majority class while underperforming on minority classes (e.g., ISUP 2+, which is the most clinically relevant but least common).

Conclusion:

Hyperparameter optimization plays a critical role in enhancing the performance of ML models. In this study, GridSearchCV was employed to systematically search for the best combination of hyperparameters, such as the number of trees, tree depth, and splitting criteria, to build a robust RF model for PCa classification. The use of K-Fold Cross-Validation (specifically the stratified version) ensures that model evaluation is consistent and fair across different folds, thereby leading to more reliable and generalized predictions. This optimization process ultimately aims to create a model that can provide accurate, reproducible predictions in the context of PCa diagnosis, contributing to better patient outcomes.

5.9. Model Training

Training Procedure:

After the hyperparameter optimization process, the RF model was trained using the best set of hyperparameters identified through GridSearchCV. This training phase involved fitting the RF model on the training dataset with the chosen combination of parameters to maximize performance and minimize errors.

The training process utilized the following steps:

- Best Hyperparameters Application: The best hyperparameters identified through GridSearchCV were applied to configure the final model. These included:
    - Number of Estimators (n_estimators): The optimal number of trees was used to balance computational cost with model performance.
    - Max Depth (max_depth): The depth of each tree was set based on the grid search results to control overfitting while maintaining the complexity required for capturing patterns in the data.
    - Minimum Samples per Split (min_samples_split) and Minimum Samples per Leaf (min_samples_leaf): These parameters ensured that each split and each leaf node had enough data to prevent overfitting to specific data points.
    - Class Weight (class_weight='balanced'): This parameter was maintained in the final model to handle class imbalance effectively, ensuring that minority classes (e.g., ISUP 2+) were not overlooked.

- Data Preparation: The training dataset was used with appropriate preprocessing steps—numerical variables were standardized where needed, and categorical variables were encoded to ensure that the RF could process the data effectively.

- Fitting the Model: The model was trained on the training set, using the best hyperparameters found. During this training process, the model constructed multiple decision trees, each using a different bootstrap sample of the data. The trees were then aggregated, with each making predictions independently, and the final prediction being determined by the majority vote of these trees.

- Evaluation During Training: The training included continuous monitoring of the model's performance using cross-validation to ensure that overfitting was minimized. The stratified K-Fold Cross-Validation approach ensured that each fold maintained a similar class distribution as the overall dataset, allowing the model to learn effectively from each class.

Results of the Optimization:

The hyperparameter tuning process using GridSearchCV provided the following best combination of hyperparameters:

- <u>Number of Estimators (n_estimators):</u> 200
  Increasing the number of estimators to 200 ensured that the model had enough trees to make stable and reliable predictions, while keeping the computational cost reasonable.

- <u>Max Depth (max_depth):</u> 20
  Setting a maximum depth of 20 helped in balancing model complexity—deep enough to capture intricate relationships in the data without causing overfitting.

- <u>Minimum Samples per Split (min_samples_split):</u> 5
  Requiring a minimum of 5 samples to split a node helped in preventing overly complex trees and ensured that each split was made with sufficient data.

- <u>Minimum Samples per Leaf (min_samples_leaf):</u> 2
  Setting a minimum of 2 samples per leaf ensured that leaf nodes contained enough data to provide stable predictions, reducing the risk of overfitting.

- <u>Maximum Features (max_features):</u> 'sqrt'
  Using the square root of the number of features at each split maintained diversity among the trees by ensuring that different splits were made based on different subsets of features.

- <u>Class Weight (class_weight):</u> 'balanced'
  This setting automatically adjusted the weights for each class based on their occurrence, effectively addressing the imbalance issue in the ISUP classes and ensuring that aggressive cancer cases (which were underrepresented) were adequately emphasized.

Conclusion:

The model training phase was conducted using the best hyperparameters determined by the optimization process, ensuring that the RF model was well-configured to handle the complexity of the dataset. By tuning key hyperparameters, the model was able to achieve better generalization, improved accuracy, and sensitivity to the underrepresented classes. These optimizations ensured that the model could provide reliable predictions for PCa classification, thereby supporting clinical decision-making.

5.10. Model Evaluation

Evaluation Metrics:

Evaluating the performance of a classification model, especially in the medical domain, requires multiple metrics to gain a comprehensive understanding of its strengths and weaknesses. The following metrics were used to assess the RF model trained for PCa classification based on ISUP grades.

1. Accuracy:

- Definition: Accuracy is the ratio of correctly predicted instances to the total number of instances. It is calculated as:

$$Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}$$

- Interpretation: Accuracy provides an overall measure of how well the model is performing across all classes. While accuracy is easy to understand and widely used, it can be misleading in cases of imbalanced datasets. In this study, since the dataset is imbalanced (e.g., fewer cases of ISUP 2+), accuracy alone may not provide a true representation of model performance, particularly for underrepresented classes.

2. Classification Report:

The classification report includes three main metrics for each class: precision, recall, and f1-score.

- Precision: Precision is the ratio of true positive predictions to the total predicted positives for a particular class. It indicates how many of the model's positive predictions were actually correct. High precision is critical in medical settings, where minimizing false positives is important.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall (Sensitivity): Recall is the ratio of true positive predictions to all actual positives for a class. It shows the model's ability to correctly identify all instances of a class. High recall is especially important for the more aggressive ISUP grades, where missing a positive case could lead to under-treatment.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- F1-Score: The F1-score is the harmonic mean of precision and recall. It is a balanced measure that is useful when there is a need to find an optimal trade-off between precision and recall, particularly in cases of imbalanced datasets.

$$F1_{Score} = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

- Interpretation: The classification report for this RF model includes precision, recall, and F1-score for each ISUP class (0, 1, 2+). By examining these metrics individually for each class, we can understand the model's effectiveness in distinguishing between less aggressive cancers (ISUP 0 and 1) and more aggressive forms (ISUP 2+).

3. Confusion Matrix:

The Confusion Matrix is a table that provides a comprehensive overview of the model's performance by comparing actual labels with predicted labels.

- Interpretation: The confusion matrix presents true positives, true negatives, false positives, and false negatives for each class. This allows us to analyze specific types of errors the model makes, such as:
  - False Negatives (FN): Especially critical in this medical application, as missing an aggressive cancer case (e.g., ISUP 2+) could have serious implications for patient care.
  - False Positives (FP): Can lead to unnecessary treatments or additional follow-up procedures for patients, increasing healthcare costs and patient anxiety.

The confusion matrix helps to identify whether the model is disproportionately misclassifying certain ISUP grades, indicating areas for potential improvement.

4. ROC-AUC Score:

- ROC Curve: The Receiver Operating Characteristic (ROC) curve is used to evaluate the model's ability to distinguish between classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. In a multiclass classification problem, ROC curves are generated for each class using a "one-vs-rest" approach.

- AUC: The Area Under the Curve (AUC) is a scalar value ranging from 0 to 1, representing the likelihood that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. An AUC closer to 1 indicates good performance, while an AUC near 0.5 suggests no discriminative power.

- Importance in Multiclass Problems: In multiclass classification, ROC-AUC helps understand the model's discriminative ability for each ISUP grade. This is important to ensure that the model not only achieves high accuracy but also maintains strong discrimination capabilities across all classes, particularly for high-risk cases like ISUP 2+.

5.11. Results

This section presents the key outcomes of the RF model developed to classify PCa patients based on ISUP grades. The focus is on model performance across multiple metrics, highlighting its ability to differentiate between absence of PCa (ISUP 0), ncsPCa (ISUP 1) and more aggressive forms- csPCa (ISUP 2+).

Model Performance Results

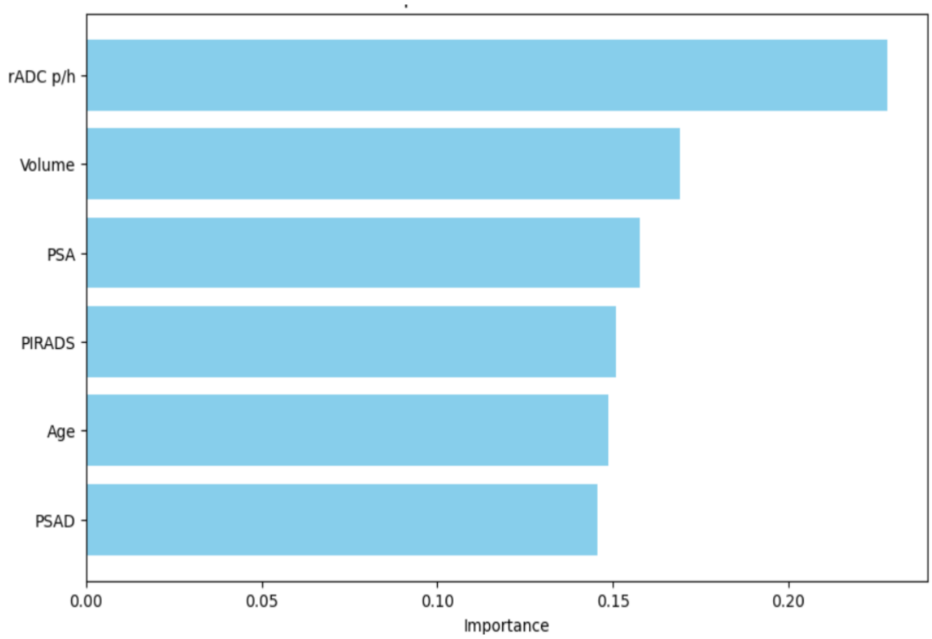
The classification model was evaluated on the test set, consisting of 95 instances across three PCa severity classes (ISUP 0, 1, and 2+). The classification report provides key insights into the model's ability to distinguish between these classes, highlighting metrics such as precision, recall, and F1-score.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
ISUP 0	0.79	0.83	0.81	59
ISUP 1	0.22	0.12	0.15	17
ISUP 2+	0.58	0.74	0.65	19
Accuracy			0.68	95
Macro Avg	0.53	0.56	0.54	95
Weighted Avg	0.65	0.68	0.66	95

Classification Report for ISUP Grades

Feature Importance Analysis

The feature importance analysis performed on the RF model highlights the relative predictive power of each variable in classifying PCa severity. The chart shows that the most influential feature is the rADC (with an importance score exceeding 0.20), followed by Prostate Volume and PSA. Other features, such as PI-RADS, Age, and PSAD, also contribute to the model, though with slightly lower importance values. This distribution of importance reflects the model's assessment of each variable's contribution to improving classification performance.

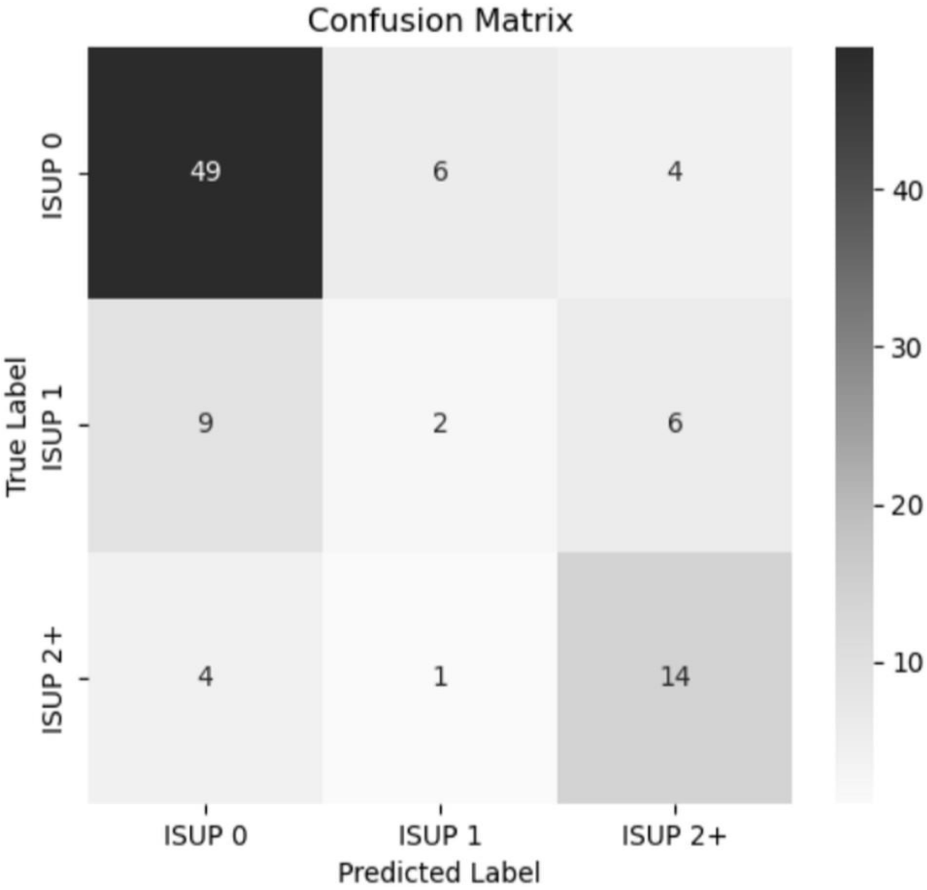


Feature importance in RF model

The feature importance metric in RF models is based on the average decrease in node impurity across all trees in the ensemble. Higher importance values indicate features that, when used for splitting, reduce impurity (e.g., the Gini impurity or entropy) significantly, enhancing the model's predictive capacity.

Confusion Matrix

The confusion matrix provides a detailed overview of the model's performance in classifying PCa severity across three classes, based on the test set. This matrix is constructed from the predictions made on the test set, which contains a total of 95 instances.

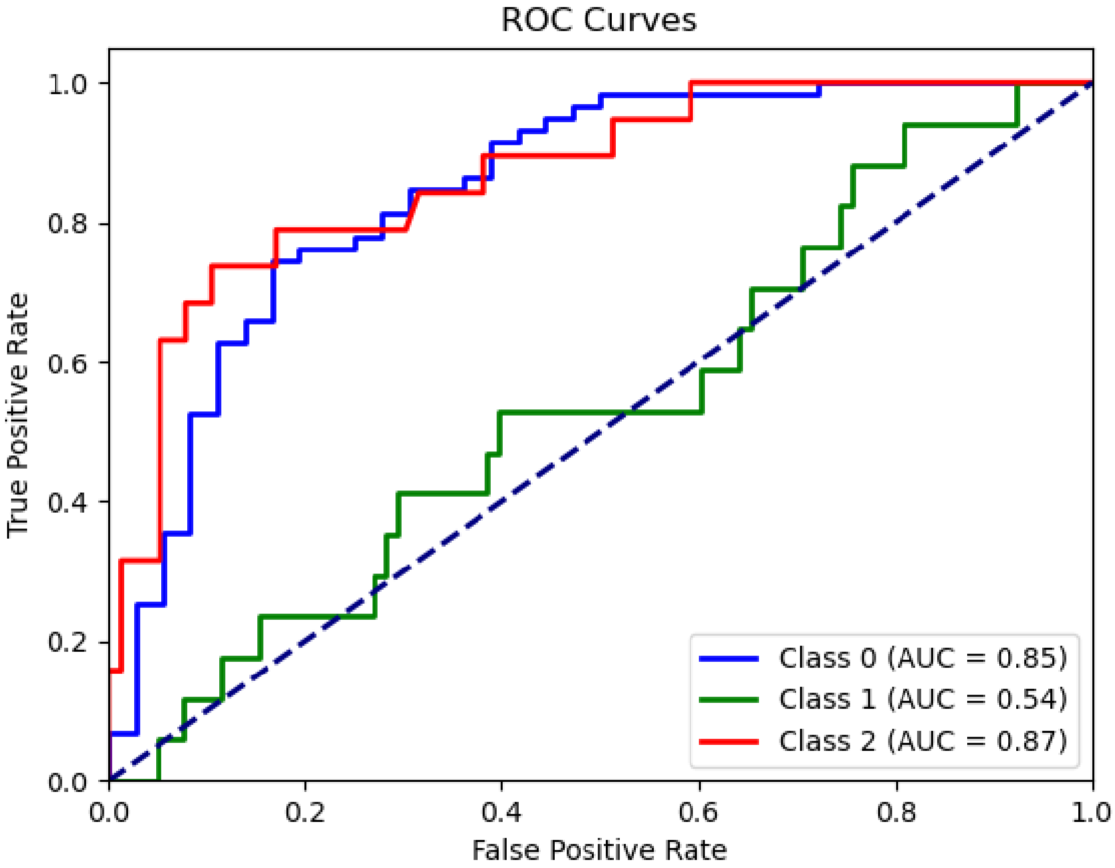


As shown in the matrix, the model demonstrates a strong ability to correctly classify instances of class 0, with 49 true positives. However, some misclassifications are observed between class 1 and class 2, where instances of class 1 are sometimes predicted as class 0 (9 instances) or class 2 (6 instances), and vice versa. Similarly, for class 2, 14 instances are correctly classified, but there are 4 misclassifications into class 0 and 1 into class 1.

These misclassifications highlight the challenge in distinguishing between neighboring ISUP grades, particularly between class 1 and class 2. Despite this, the confusion matrix confirms the model's overall effectiveness on the test set, offering key insights into areas where additional feature engineering or model refinement could further improve classification accuracy.

ROC Curves for ISUP Classification

To evaluate the RF model’s classification performance across different ISUP grades, ROC curves were generated using a "one-vs-rest" approach for each class (ISUP 0, 1, and 2+). This approach allowed us to assess the model’s ability to discriminate between classes by focusing on the True Positive Rate (TPR) and False Positive Rate (FPR) for each ISUP grade [10 -11-12-13-14-15-16-17-18-19-20-21-22-23-24-25-26-27-28-29].


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.85)
Class 1 (AUC = 0.54)
Class 2 (AUC = 0.87)

5.12. Discussion

Model Performance Overview

The RF classifier developed for classifying PCa severity based on ISUP grades (0, 1, 2+) was evaluated using precision, recall, F1-score, and overall accuracy. The test set included 95 cases across three ISUP classes, reflecting varying cancer severity levels.

- ISUP 0 (No Cancer): The model achieved strong performance in identifying cases without cancer, with precision, recall, and F1-score values of 0.79, 0.83, and 0.81, respectively. This high performance reduces the risk of unnecessary interventions by accurately distinguishing non-cancerous cases.

- ISUP 1 (ncsPCa): Performance for ISUP 1, was lower, with precision and recall at 0.22 and 0.12, respectively. The model's limited ability to differentiate these cases highlights the challenges of distinguishing this category from both negative and high-risk cases. This limitation suggests a need for improved feature engineering or alternative modeling approaches.

- ISUP 2+ (csPCa): For this group, the model reached a precision of 0.58 and a recall of 0.74. This moderate performance reflects a reasonable capacity to identify aggressive cases, though some high-severity instances were misclassified as lower grades. Recognizing such cases is critical for ensuring timely treatment of high-risk patients.

The model demonstrates robust classification capability for ISUP 0 cases, while performance for ISUP 2+ cases is reasonable but leaves room for improvement. However, for ISUP 1, the model's suboptimal performance reflects the inherent limitations of prostate MRI, which historically underestimates or misses low-risk, ncsPCa. Since our model relies on categorical data from MRI (PI-RADS and rADC), rather than advanced MRI image analysis techniques, it is particularly susceptible to this bias. Consequently, if MRI fails to detect ISUP 1 lesions, the model is unlikely to accurately classify these cases. Integrating direct MRI image analysis using dedicated software in future iterations is expected to significantly enhance detection of low-grade lesions.

Overall, the model achieves an accuracy of 0.68 on the test set, with a macro-average F1-score of 0.54, indicating moderate balance across all classes, and a weighted-average F1-score of 0.66, reflecting stronger performance in the more prevalent ISUP 0 class. These results confirm the model's capability in detecting significant patterns while identifying areas for improvement, particularly in classifying ISUP 1 cases.

Feature Importance interpretation

The feature importance results reveal some counterintuitive aspects, such as the relatively lower role of PI-RADS compared to variables like rADC or PSA. This can be explained by several factors related to the model's nature and the dataset structure. Firstly, some variables may be redundant, meaning they provide similar information. When the model finds similar information in multiple features, it tends to reduce the importance of a specific variable, favoring those with higher informational variability.

Additionally, derived variables like PSAD (which normalizes PSA relative to prostate volume) may cause a reduction in the importance of the original variables it derives from. This occurs because the model can find a more synthetic and linear representation of information through composite variables, reducing the need to rely on the base components.

Finally, it’s worth noting that the feature importance in a RF model reflects impurity reduction in tree splits, which doesn’t necessarily align with the direct clinical relevance of variables.

Confusion Matrix Analysis

The confusion matrix provides further insight into the classification strengths and weaknesses:

- **True Positives:** The model demonstrated effective classification for ISUP 0 cases with 49 true positives. This strengthens confidence in the model’s reliability for no-cancer predictions.

- **Misclassifications:** Misclassifications were most common between ISUP 1 and ISUP 2+, suggesting that the model struggles with the subtle clinical differences between these adjacent grades. These errors indicate areas for improvement, such as the inclusion of additional discriminative features.
    - <u>False Negatives (ISUP 2+):</u> Misclassifications where ISUP 2+ cases are labeled as lower grades pose a clinical risk as these patients may receive delayed treatment. These errors suggest a need for more refined features or additional data to improve model sensitivity for high-grade cancers.
    - <u>False Positives: </u>Cases misclassified as higher grades, particularly from ISUP 0 to ISUP 1 or 2+, could lead to overtreatment. While less concerning than false negatives, false positives still have implications for patient care, including unnecessary interventions.

ROC Curve and AUC Analysis

ROC curve evaluates the model's performance in distinguishing between different PCa severity categories, as defined by ISUP grading:

- **Class 0 (ISUP 0)**: Have an AUC of 0.85. This high AUC indicates that the model effectively differentiates non-cancerous cases from cancerous ones, showing strong discriminatory power for identifying patients without tumors.

- **Class 1 (ISUP 1)**: Have an AUC of 0.54. This low AUC reveals that the model struggles to distinguish these low-risk cases from other categories, demonstrating limited sensitivity and specificity for ISUP 1. The difficulty in classifying ISUP 1 cases may stem from the subtle features of ncsPCa, which are often underestimated or missed by MRI due to its intrinsic limitations.

- **Class 2 (ISUP 2+)**: Have an AUC of 0.87, suggesting high accuracy in identifying high-risk cases. The model performs well in distinguishing csPCa, likely due to the more pronounced features associated with these tumors.

The model demonstrates strong performance in identifying the extremes: patients without cancer (ISUP 0) and those with csPCa (ISUP 2+). However, the limited performance for ISUP 1 cases (AUC 0.54) indicates a need for improvement, as these low-risk, ncsPCa are challenging to classify due to their subtle characteristics. Enhancing model performance for ISUP 1 detection could involve incorporating advanced imaging analysis techniques or additional data features to mitigate MRI's inherent bias in detecting ncsPCa.

Future Work

The RF classifier demonstrates considerable potential as a predictive tool for assessing PCa severity, particularly in identifying no-cancer (ISUP 0) and high-risk (ISUP 2+) cases. However, its limitations in classifying ISUP 1 cases highlight areas for further refinement. Moving forward, several improvements are recommended to enhance the model's accuracy, interpretability, and clinical utility.

- **Feature Enhancement**: Expanding the feature set to include additional radiomic and clinical variables could enhance the model's ability to differentiate between ISUP 1 and ISUP 2+ cases. The integration of imaging biomarkers through the FLUTE project, using

QUIBIM’s QP Prostate software, will enable the inclusion of advanced radiomic features that capture tumor microarchitecture and biological characteristics. This addition is expected to improve the model’s sensitivity for detecting csPCa in ISUP 2+ cases by offering a more granular view of tumor heterogeneity and progression potential.

- **Addressing Class Imbalance**: To improve the model’s recall for underrepresented ISUP grades, particularly ISUP 2+, future work should explore oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) or hybrid methods that combine oversampling and undersampling. This approach would create a more balanced dataset, enabling the model to better identify high-risk cases and reduce the likelihood of under-diagnosing aggressive cancers.

- **Advanced Model Tuning and Ensemble Approaches**: Exploring ensemble methods could further enhance the model’s sensitivity and specificity, especially for ISUP 2+ cases. Incorporating boosting techniques, such as XGBoost, alongside RF may improve overall predictive performance. A hybrid approach—using RF to capture complex feature interactions and XGBoost to sequentially reduce error rates—could enhance classification accuracy across all ISUP grades.

- **Enhancing Interpretability**: For effective clinical adoption, model interpretability is crucial. Applying interpretability methods such as SHAP (SHapley Additive exPlanations) can provide local, model-agnostic explanations, clarifying feature contributions for individual predictions. This transparency will build trust among clinicians and help identify factors contributing to high-risk predictions, ensuring alignment with clinical understanding.

- **Data Expansion**: Expanding the dataset, particularly for ISUP 2+ cases, would increase the model’s generalizability and robustness. Collaborating with additional healthcare institutions or accessing publicly available datasets could provide a more comprehensive data source. Longitudinal data collection would further support predictive modeling by capturing progression patterns over time, enhancing the model's ability to predict higher ISUP grades.

- **Developing a Robust Clinical Tool**: The long-term goal is to develop a clinically viable tool that is accurate, interpretable, and suitable for real-time use. Future refinements will focus on balancing predictive power with usability across diverse clinical settings. Using FL approaches through the FLUTE project will facilitate this by training on multi-institutional data. This collaborative approach will yield a more generalizable model, adapting to regional variations and making it a valuable tool for broader clinical use.

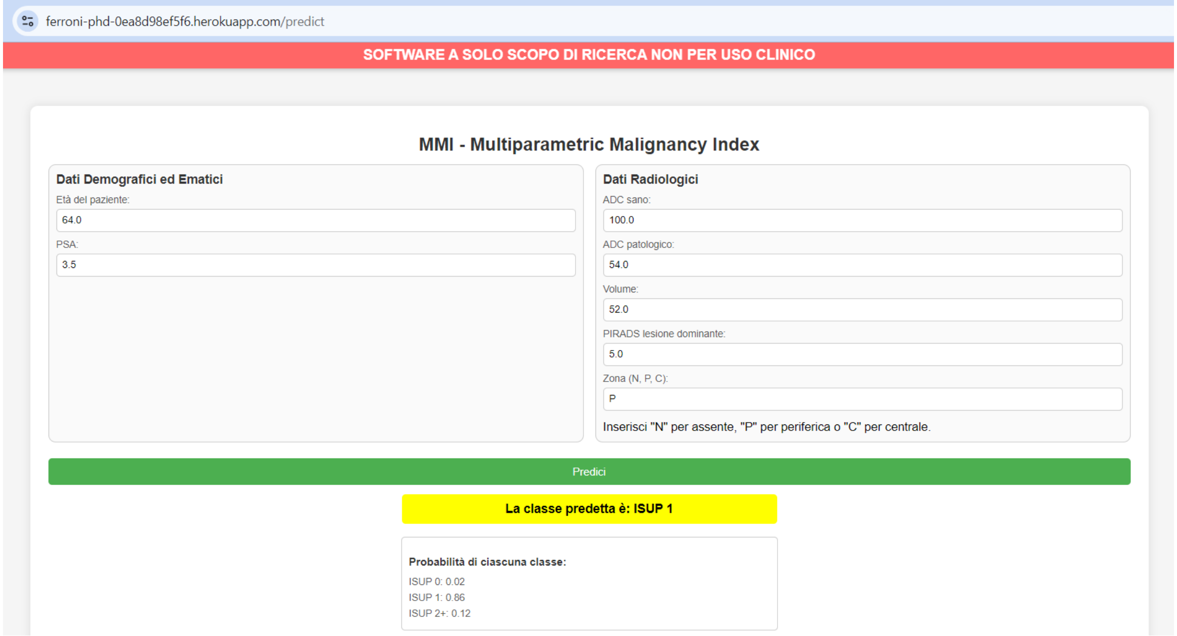
In summary, these future directions aim to refine the RF model's accuracy, interpretability, and clinical applicability. Addressing class imbalance, improving model transparency, exploring ensemble techniques, expanding the dataset, and integrating advanced imaging biomarkers will help evolve the model into a highly reliable tool for predicting PCa severity. These advancements align with the goals of the FLUTE project, supporting a shift towards precision medicine in PCa care, where AI-driven tools play an essential role in personalized diagnosis and treatment planning.

6. Development of a Web Application for MMI evaluation

In this chapter, we describe the development of a web application that integrates a machine learning model for predictive purposes. The project aimed to create a user-friendly interface for non-technical users to input clinical data and obtain real-time predictions based on a pre-trained Random Forest (RF) model. After the training process, the model was serialized and exported as a .pkl file using joblib, allowing it to be seamlessly integrated into a web-based application.

By deploying this trained model on a cloud-hosted platform like Google Cloud, it became possible to calculate risk classes for new patient data entered through a web interface. This solution enables users, such as clinicians and researchers, to input new clinical parameters via the portal and instantly receive predictions of the ISUP class.

The solution is built using Python for backend processing and seamless integration of the ML model. The application is served in a robust production environment through Gunicorn, while Google Cloud App Engine is used as the cloud deployment platform, making the application easily accessible to users online. This setup leverages the scalability and reliability of Google Cloud, ensuring that the application can handle multiple user requests and provide consistent performance.



ferroni-phd-0ea8d98ef5f6.herokuapp.com/predict

SOFTWARE A SOLO SCOPO DI RICERCA NON PER USO CLINICO

MMI - Multiparametric Malignancy Index

Dati Demografici ed Ematici

Età del paziente: 64.0

PSA: 3.5

Dati Radiologici

ADC sano: 100.0

ADC patologico: 54.0

Volume: 52.0

PIRADS lesione dominante: 5.0

Zona (N, P, C): P

Inserisci "N" per assente, "P" per periferica o "C" per centrale.

La classe predetta è: ISUP 1

Probabilità di ciascuna classe:

ISUP 0: 0.02
ISUP 1: 0.86
ISUP 2+: 0.12

Multiparametric Malignancy Index (MMI) Web Interface. *The figure shows the web-based tool designed to predict ISUP grades using both clinical and radiological data. Users input demographic, hematologic, and radiological features, including patient age, PSA level, and ADC values, to receive a real-time ISUP grade prediction for PCa severity. The predicted class and the probability of each ISUP category are displayed for interpretation.*

This web-based approach serves as an example of how a trained model can be made accessible globally, allowing others to predict risk for their own patients. It demonstrates that real-time risk assessment can be performed remotely, without the need for local resources or ML expertise.

https://ferroni-phd.oa.r.appspot.com/

7. Conclusions

The RF model demonstrated moderate potential in classifying PCa severity, showing reasonable discriminatory power, especially in distinguishing between ISUP 0 (no cancer) and ISUP 2+ cases (csPCa). However, the model's performance for ISUP 1 cases (ncsPCa) was considerably lower, as reflected by moderate AUC scores in the ROC curve analysis: 0.85 for ISUP 0, 0.87 for ISUP 2+, and only 0.54 for ISUP 1. These findings suggest that, while the model effectively identifies the absence of cancer and highly aggressive cases, it lacks sensitivity in detecting ncsPCa (ISUP 1), highlighting a crucial area for improvement.

In particular, the RF model's limited sensitivity for ISUP 1 cases, suggests a need for enhanced model tuning or further data balancing to better capture these cases. The lower AUC for ISUP 1 indicates that future model iterations should focus on improving recall and precision for this group to ensure accurate classification. Additionally, this underperformance may not only reflect the need for additional features or model adjustments but also highlight the inherent limitations of prostate MRI, which historically tends to underestimate ncsPCa.

The feature importance analysis identified rADC, Volume, and PSA as the most influential predictors, with rADC having the highest importance score. These findings underscore the clinical relevance of these features, yet indicate that PI-RADS and PSAD, while included, play a lesser role in the current model's predictions. Although these variables are well-established in clinical practice, their relatively lower importance in this model suggests that incorporating additional predictive features could improve accuracy, especially for intermediate-risk cases like ISUP 1. This highlights the potential benefit of enhancing the model with supplementary data to address the complexity of these cases more effectively.

This study provides a foundational framework for stratifying PCa severity using RF models based on clinical and radiological data. The model achieved satisfactory classification for ISUP 0 and ISUP 2+, indicating that it can be a valuable tool in specific clinical scenarios. However, the results underscore the need for further refinement, particularly in addressing the limitations seen in intermediate-risk classifications (ISUP 1).

Looking ahead, this work sets a foundational stage for the FLUTE project, which aims to advance PCa diagnosis through integrating AI and FL. Future model iterations will incorporate quantitative imaging features from multiparametric MRI using QP Prostate software, developed by QUIBIM. By adding these advanced radiomic features, we aim to achieve a more nuanced representation of tumor biology and microarchitecture, potentially enhancing the model's accuracy across all ISUP grades.

In the words of Hippocrates, "Cure sometimes, treat often, comfort always".' This AI-driven journey in PCa classification stands as a testament to precision medicine's promise: to transform complexity into clarity, guiding clinicians to make informed, personalized decisions. By advancing these tools, we step closer to a future where diagnostics seamlessly integrate data and expertise, providing hope and precision for every patient.

In an era where medicine intertwines with AI, every piece of data becomes an opportunity to save lives and elevate patient care. This explorative model is not just a step toward more accurate diagnoses, but a leap toward a future where technology enhances the human touch. As William Osler said, 'Medicine is a science of uncertainty and an art of probability.' With these new frontiers, we turn uncertainty into knowledge for a better tomorrow.

8. ANNEXES

8.1. Python Code for RF Model

The following annex provides the complete Python code used for the implementation, training, and evaluation of the RF model for the classification of PCa based on clinical and radiological data. The code is divided into several key stages, including data preprocessing, model training, hyperparameter tuning, and evaluation using performance metrics.

python CODE:

Part 0: Libreries

```
import pandas as pd
import time  # For tracking training time
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
```

Part 1: Data Loading and Splitting

```
# Load and split data into training and test sets
def load_and_split_data(file_path):
    data = pd.read_csv(file_path)
    X = data.drop(columns=['isup_classi (0, 1, 2)'])
    y = data['isup_classi (0, 1, 2)']
    return train_test_split(X, y, test_size=0.3, random_state=42)
```

```
# Load data from file
file_path = r'C:\PHD\Prostate_DB.csv'
X_train, X_test, y_train, y_test = load_and_split_data(file_path)
```

Part 2: Preprocessing and Model Creation

```
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier


# Create the model with a preprocessing pipeline
def create_model(X_train):
    numeric_transformer = Pipeline(steps=[('scaler', StandardScaler())])
    preprocessor = ColumnTransformer(transformers=[
        ('num', numeric_transformer, X_train.select_dtypes(include=['float64',
'int64']).columns)
    ])
    random_forest = RandomForestClassifier(class_weight='balanced',
random_state=42)
    pipeline = Pipeline(steps=[('preprocessor', preprocessor), ('classifier',
random_forest)])
    return pipeline


# Create the pipeline model
pipeline = create_model(X_train)
```

Part 3: Hyperparameter Tuning with GridSearchCV

```
from sklearn.model_selection import GridSearchCV

import time



# Perform hyperparameter tuning using GridSearchCV

def perform_hyperparameter_tuning(pipeline, X_train, y_train):

    param_grid = {

        'classifier__n_estimators': [100, 200, 300],

        'classifier__max_depth': [10, 20, 30, None],

        'classifier__min_samples_split': [2, 5, 10],

        'classifier__min_samples_leaf': [1, 2, 4],

        'classifier__max_features': ['sqrt', 'log2']

    }

    grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='accuracy',
n_jobs=-1)



    start_time = time.time()

    grid_search.fit(X_train, y_train)

    elapsed_time = time.time() - start_time

    print(f"Training complete. Time taken: {elapsed_time:.2f} seconds.")



    return grid_search.best_estimator_



# Run hyperparameter tuning

best_model = perform_hyperparameter_tuning(pipeline, X_train, y_train)
```

References

1. EAU - EANM - ESTRO - ESUR - ISUP - SIOG Guidelines on Prostate Cancer; P. Cornford, D. Tilki, R.C.N. van den Bergh, Et all (https://uroweb.org/guidelines/prostate-cancer)
2. PI-RADS V2.1 ACR (https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/PI- RADS)
3. Big data, artificial intelligence, and structured reporting - Daniel Pinto dos Santos and Bettina Baeßler - European Radiology Experimental (2018) 2:42 https://doi.org/10.1186/s41747-018-0071- 4
4. https://www.fluteproject.eu
5. https://www.mripro.io
6. Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563-577.
7. Cuocolo, R., et al. (2021). Radiomics and Magnetic Resonance Imaging of Prostate Cancer: A Systematic Review and Radiomic Quality Score Assessment. European Urology Oncology, 4(3), 429-432.
8. Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.
9. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.
10. Breiman, L. "Random forests." Machine learning 45.1 (2001): 5-32.
11. Dietterich, T. G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000.
12. Hastie, T., Tibshirani, R., & Friedman, J. "The Elements of Statistical Learning." Springer Series in Statistics (2009).
13. Kuhn, M., & Johnson, K. "Applied Predictive Modeling." Springer Science & Business Media, 2013.
14. Courtiol, P., et al. "Deep learning-based classification of tumor grade." Nature Communications 10 (2019): 1-10.
15. Rahman, M. M., et al. "Handling class imbalance in prostate cancer data using resampling techniques." Bioinformatics 36.22 (2020): 5321-5327.
16. Chen, T., & Guestrin, C. "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD (2016).
17. Pal, M. "Random forest classifier for remote sensing classification." International Journal of Remote Sensing 26.1 (2005): 217-222.
18. Han, J., Kamber, M., & Pei, J. "Data Mining: Concepts and Techniques." Morgan Kaufmann, 2011.
19. Pedregosa, F., et al. "Scikit-learn: Machine learning in Python." JMLR 12 (2011): 2825-2830.
20. Lundberg, S. M., & Lee, S. I. "A unified approach to interpreting model predictions." NIPS, 2017.
21. Van Ginneken, B., et al. "Comparison of logistic regression, k-nearest neighbor, and SVM classifiers in lung nodule detection." Medical Imaging 22.5 (2002): 653-663.
22. Guyon, I., & Elisseeff, A. "An introduction to variable and feature selection." JMLR 3 (2003): 1157- 1182.
23. Sandri, M., & Zuccolotto, P. "Variable selection using random forests." Data Analysis (2008): 263- 270.
24. Esteva, A., et al. "A guide to deep learning in healthcare." Nature Medicine 25 (2019): 24-29.
25. Quinlan, J. R. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.
26. Varma, S., & Simon, R. "Bias in error estimation when using cross-validation for model selection." BMC bioinformatics 7.1 (2006): 1-8.
27. Criminisi, A., Shotton, J., & Konukoglu, E. "Decision forests: A unified framework for classification, regression, density estimation, manifold learning, and semi-supervised learning." Foundations and Trends® in Computer Graphics and Vision 7.2–3 (2012): 81-227.
28. Lu, M. Y., et al. "Data-efficient and weakly supervised computational pathology." Nature Biomedical Engineering 5.6 (2021): 555-570.
29. Pölsterl, S. "Handling missing values in machine learning." Advances in Data Science (2018): 105- 124.