TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

DATA ANALYSIS WITH LIMITED DATA AVAILABILITY

Prostate Cancer Prediction and Characterization as a Case Study

Ileana Montoya Perez

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 36 | TECHNICA – INFORMATICA | TURKU 2024

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

DATA ANALYSIS WITH LIMITED DATA AVAILABILITY

Prostate Cancer Prediction and Characterization as a Case Study

Ileana Montoya Perez

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 36 | TECHNICA – INFORMATICA | TURKU 2024

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral programme in Technology (DPT)

Supervised by

Research Director
Professor, Jukka Heikkonen
Department of Computing
University of Turku, Finland

Professor, Tapio Pahikkala
Department of Computing
University of Turku, Finland

Associate Professor, Peter J. Boström
Department of Urology
University of Turku
Turku University Hospital, Finland

Adjunct Professor, Ivan Jambor
Department of Radiology
University of Turku
Turku University Hospital, Finland

Reviewed by

Professor, Hanna Suominen
College of Engineering,
Computing and Cybernetics
The Australian National University

Associate Professor, Ilkka Pölönen
Computational Data Analysis
Faculty of Information Technology
University of Jyväskylä, Finland

Opponent

Professor, Sébastien Lafond
Department of Information Technologies
Faculty of Science and Engineering
Åbo Akademi University, Finland

ISBN 978-951-29-9645-2 (PRINT)
ISBN 978-951-29-9646-9 (PDF)
ISSN 2736-9390 (PRINT)
ISSN 2736-9684 (ONLINE)
Painosalama, Turku, Finland, 2024

To my family

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
MONTOYA PEREZ, ILEANA: Data analysis with limited data availability
Doctoral dissertation, 144 pp.
Doctoral programme in Technology (DPT)
May 2024

ABSTRACT

Research studies conducted on limited datasets (i.e., data from tens to maximum hundreds of observations) may be the only practical option for many research areas, as data collection might be costly, complex, or both. Data analysis on these datasets is challenging as it can lead to inaccurate results. In this thesis, we addressed this challenge in the context of prostate cancer research by empirically assessing the predictive and characterization capabilities of attributes with the following objectives: to evaluate the predictive power of features extracted from prostate magnetic resonance imaging (MRI) using cross-validation techniques, to develop and evaluate a cross-validation method for small sample sizes that allow receiver operating characteristic (ROC) analysis, and to identify and compare relevant predictors among MRI features, clinical variables, gene expressions, and kallikreins for prostate cancer detection and stratification. To achieve these objectives, we used data from approved studies and registered clinical trials at Turku University Hospital, involving a strong collaboration between university departments and hospitals. This collaboration enabled the collection of diverse, high-quality features to enhance prostate cancer diagnosis and prognosis research.

The results of this thesis can be summarized as follows. First, when evaluating radiomic features from various MRI modalities, our findings demonstrate the potential that these features have in stratifying prostate tumors into low- and high-risk. Second, in terms of model evaluation using ROC analysis and cross-validation, our research highlights a significant negative bias in the area under the ROC curve when estimated by leave-one-out (LOOCV) and introduces a novel cross-validation method called tournament leave-pair-out (TLPOCV) as a more reliable method for ROC analysis than LOOCV. Finally, our results provide empirical evidence of the predictive potential that quantitative and qualitative features from MRI, clinical variables, gene expressions, and kallikreins—individually and in combination—have in detecting and stratifying prostate cancer.

The findings in this research are of interest not only to medical professionals and healthcare providers engaged in prostate cancer research but also to those involved in analyzing and learning from size-constrained datasets while achieving clinically meaningful evaluation outcomes.

KEYWORDS: prostate cancer, data analysis, small sample, model evaluation, cross-validation, ROC curve

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Oppiaine
MONTOYA PEREZ, ILEANA: Data analysis with limited data availability
Väitöskirja, 144 s.
Tohtoriohjelma
Toukokuu 2024

TIIVISTELMÄ

Rajallisilla aineistoilla suoritettavat tutkimukset (kymmenistä satoihin havaintoihin) ovat joskus ainoa käytännöllinen vaihtoehto aineiston keräämisen kalleuden, monimutkaisuuden tai molempien vuoksi. Pieni koko taas voi johtaa epätarkkoihin tuloksiin. Tässä väitöskirjassa käsittelemme tätä haastetta eturauhassyöpätutkimuksen kontekstissa arvioimalla empiirisesti attribuuttien ennustavaa ja karakterisointikykyä seuraavilla tavoitteilla: arvioida eturauhasen magneettikuvauksesta (MRI) saaduista piirteistä ekstraktoitujen ominaisuuksien ennustuskykyä ristiinvalidointitekniikoilla, kehittää ja arvioida ristiinvalidointimenetelmä ROC-analyysin mahdollistamiseksi pienillä näyteko'oilla sekä tunnistaa ja vertailla merkityksellisiä ennustajia MRI-piirteiden, kliinisten muuttujien, geeniekspressioiden ja kallikreiinien välillä eturauhassyövän havaitsemiseksi ja stratifioimiseksi. Näiden tavoitteiden saavuttamiseksi käytimme tietoja hyväksytyistä tutkimuksista ja rekisteröidyistä kliinisistä kokeista Turun yliopistollisessa keskussairaalassa. Tämä yhteistyö mahdollisti monipuolisten ja laadukkaiden piirteiden keräämisen eturauhassyövän diagnoosin ja ennusteen tutkimuksen parantamiseksi.

Tämän väitöskirjan tulokset voidaan tiivistää seuraavasti. Löydöksemme vahvistivat radiomisten piirteiden potentiaalin stratifioida eturauhaskasvaimia matala- ja korkeariskisiin. Toiseksi, tutkimuksemme osoitti yksittäisristiinvalidointiin perustuvien estimaattorien olevan negatiivisesti harhautuneita erottelukykykäyrän (ROC-käyrä) alaisen pinta-alan (AUC) arvioinnissa, ja esittelee uuden ristiinvalidointimenetelmän nimeltä turnajaisyksittäisristiinvalidointi (TLPOCV), joka välttää kyseisen harhan. Lopuksi tuloksemme tarjoavat empiiristä näyttöä siitä ennustepotentiaalista, joka kvantitatiivisilla ja kvalitatiivisilla piirteillä MRI:stä, kliinisillä muuttujilla, geeniekspressioilla ja kallikreiineillä on eturauhassyövän havaitsemisessa ja stratifioimisessa sekä yksittäin että yhdistettynä. Tämän tutkimuksen löydökset ovat hyödyllisiä paitsi eturauhassyöpätutkimuksessa mukana oleville lääketieteen ammattilaisille, myös yleisemmin pienikokoisia datoja analysoiville tutkijoille.

ASIASANAT: eturauhassyöpä, aineistoanalyysi, pieni otos, mallin ennustuskyvyn arviointi, ristiinvalidointi, ROC-käyrä

Acknowledgements

I am truly privileged to have pursued my doctoral studies in a collaborative environment that involved joint efforts between university and hospital departments. This collaboration has not only enriched my academic journey with valuable resources but has also provided opportunities to work with extraordinary people, enhancing my research experience.

Firstly, I express my deepest gratitude to Professor Tapio Pahikkala, who has provided invaluable support and guidance throughout the entire journey of this doctoral research. His expertise and insightful feedback have helped shape the direction and quality of this dissertation. I also extend my thanks to my supervisors and co-authors, Associate Professor Peter Boström and Adjunct Professor Ivan Jambor. There are not enough words to express my gratitude for their support and the knowledge they have shared with me over these years. I consider myself fortunate to have them all as my supervisors and to have had the opportunity to work with them.

Throughout this journey, I have been blessed with support and guidance from many people to whom I owe gratitude. Special thanks to Professor Hannu Aronen, who supported and valued my work right from the beginning, and to Professor Jukka Heikkonen, who assisted in the completion of this dissertation. Sincere appreciation goes to Associate Professor Antti Airola, who has consistently been there to offer advice and has taken the time to provide clear explanations when I needed.

I extend my gratitude to my co-authors and colleagues: Harri Merisaari, Parisa Movahedi, and Jussi Toivonen, for their valuable contributions and advice. Additionally, I acknowledge and thank the urologists, pathologists, radiologists, and biotechnicians: Otto Ettala, Pekka Taimen, Jane Verho, Aida Steiner, Kim Pettersson, Henna Kekki, and Ferdhos Khan, for their contribution to our fluent and successful collaboration and for providing me with the opportunity to learn from their expertise.

I also thank Professor Hanna Suominen and Associate Professor Ilkka Pölönen for acting as pre-examiners and providing their valuable feedback. Furthermore, I sincerely thank Professor Sébastien Lafond for agreeing to act as my opponent.

During these years, I have been fortunate to develop work relations and friendships that have helped me navigate the ups and downs of this journey. Therefore, a big thank you goes to Jonne Pohjankukka, Petra Virjonen, Riikka Numminen, Alaleh Maskooki, Paavo Nevalainen, Anne-Maarit Majanoja, Adrian Borzyszkowski, Valtteri Nieminen, Katariina Perkonoja, and Luca Zelioli. I extend a special thank you

to Elise Haulivuori for her friendship and the incredible connection we share, for offering guidance, understanding, and a listening ear.

My deepest gratitude to my husband, Jari Komulainen. His love and belief in me have given me the strength to continue in those moments when I felt down. Also, he has provided me with a lovely Finnish family which has been a great support.

Thanks to my mom, sisters, and family. Despite the distance, we have remained close, and you have always been with me in my heart.

May 17, 2024
Ileana Montoya Perez

Table of Contents

Acknowledgements . . . . . **vi**

Table of Contents . . . . . **viii**

Abbreviations . . . . . **x**

List of Original Publications . . . . . **xiii**

1 Introduction **1**
- 1.1 Data analysis with a limited amount of data 1
- 1.2 Motivation of the research 3
- 1.3 Main objectives and research questions 4
- 1.4 Structure of the thesis 5

2 Prostate cancer domain-specifics **7**
- 2.1 Prostate gland and its anatomy 7
- 2.2 Incidence and mortality of prostate cancer 7
- 2.3 Grading of prostate cancer 8
- 2.4 Screening and Diagnosis of prostate cancer 8
- 2.5 MRI in prostate cancer diagnostics 9
  - 2.5.1 Anatomic MRI of the prostate 10
  - 2.5.2 Diffusion-weighted imaging (DWI) of prostate 11
- 2.6 Biomarkers for prostate cancer 12
- 2.7 Available datasets 12

3 Data analysis **14**
- 3.1 Statistics and machine learning 14
- 3.2 Training a model 15
  - 3.2.1 Feature selection 18
  - 3.2.2 Regularization 18
- 3.3 Evaluating a model 19
  - 3.3.1 Quantitative metrics 20
  - 3.3.2 Receiver operative characteristic (ROC) analysis . . 20
- 3.4 Selecting a model 22

3.5 Resampling techniques 23
3.5.1 Cross-validation 23
3.5.2 Tournament cross-validation for ROC analysis . . . 25
3.5.3 Nested resampling 26
3.5.4 Permutation tests 27

4 Research studies and results 29
4.1 Summary of the publications 29
4.1.1 Publication I 29
4.1.2 Publication II 31
4.1.3 Publication III 34
4.1.4 Publication IV 37
4.1.5 Publication V 40
4.1.6 Publication VI 43
4.2 Research results 45

5 Conclusions 48
5.1 Summary of the thesis 48
5.2 Discussion and outcomes 49
5.3 Future work 50

List of References 52

Original Publications 57

Abbreviations

ADC	Apparent diffusion coefficient
AUC	Area under the curve
BPH	Benign prostatic hyperplasia
CV	Cross-validation
CZ	Central zone
dPSA	PSA density
DCE	Dynamic contrast enhancement
DRE	Digital rectal examination
DWI	Diffusion-weighted imaging
EAU	European Association of Urology
FN	False negative
FP	False positive
FPR	False positive rate
GS	Gleason score
WG	Whole prostate gland
GGG	Gleason grade group
GLCM	Gray-level co-occurrence matrix
IID	Independent and identically distributed
ISUP	International Society of Urological Pathology
KNN	K-nearest neighbors learning algorithm
LBP	Local binary patterns

LOOCV	Leave-one-out cross-validation
LPOCV	Leave-pair-out cross-validation
LSOCV	Leave-subject-out cross-validation
ML	Machine learning
MRI	Magnetic resonance imaging
MSE	Mean squared error
bpMRI	Biparametric magnetic resonance imaging
mpMRI	Multiparametric magnetic resonance imaging
PZ	Peripheral zone
PCa	Prostate cancer
PSA	Prostate-specific antigen
PI-RADS	Prostate imaging reporting and data system
RALP	Robotic-assisted laparoscopic prostatectomy
RLS	Regularized least-squares
ROC	Receiver operating characteristic
SSE	Sum of squared errors
SPCa	Clinically significant prostate cancer
TN	True negative
TP	True positive
TZ	Transition zone
T2W	T2-weighted imaging
TPR	True positive rate
TYKS	Turku University Central Hospital
TLPOCV	Tournament leave-pair-out cross-validation
WMW	Wilcoxon-Mann-Whitney test

b	Diffusion weighting strength known as b-value
S_0	The signal intensity at $b = 0$
$S(b)$	Signal intensity at a particular b-value
ADC_m	Diffusion coefficient of the monoexponential function
ADC_k	Diffusion coefficient of the kurtosis function
K	Kurtosis term
$\mathcal{D}$	Data space
$\mathbf{x}$	Vector of inputs
y	Output value
$\mathcal{X}$	Set of input vectors
$\mathcal{Y}$	Set of output values
$\mathcal{A}$	Learning algorithm
$\mathcal{H}$	Hypothesis set
g	The unknown target function
f	Model that approximate g
θ	Parameter vector
$\hat{\boldsymbol{\theta}}$	Optimal parameter vector
$\mathbf{X}$	Matrix of inputs vectors
$\mathbf{y}$	Vector of output values
λ	Regularization parameter
H	Heaviside step function
$A(f)$	True unknown AUC of a function
$\hat{A}(f)$	AUC estimate of a function
$\hat{A}_{LPOCV}(f)$	Leave-pair-out cross-validation AUC estimate of a function
ζ	Coefficient of consistency
X	Random variable
π	A permutation of indices

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Ileana Montoya Perez, Jussi Toivonen, Parisa Movahedi, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J. Boström, Aida Kiviniemi, Hannu J. Aronen, Tapio Pahikkala and Ivan Jambor. Diffusion Weighted Imaging of Prostate Cancer: Prediction of Cancer using Texture Features from Parametric Maps of the Monoexponential and Kurtosis functions. 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2016; IEEE: 1-6.

II Jussi Toivonen, Ileana Montoya Perez, Parisa Movahedi, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J. Boström, Jonne Pohjankukka, Aida Kiviniemi, Tapio Pahikkala, Hannu J. Aronen and Ivan Jambor. Radiomics and Machine Learning of Multisequence Multiparametric Prostate MRI: Towards Improved Non-invasive Prostate Cancer Characterization. PLoS One, 2019; 14(7).

III Ileana Montoya Perez, Antti Airola, Peter J. Boström, Ivan Jambor and Tapio Pahikkala. Tournament Leave-Pair-Out Cross-validation for Receiver Operating Characteristic Analysis. Statistical Methods in Medical Research, 2019; 28(10-11): 2975-2991.

IV Ileana Montoya Perez, Ivan Jambor, Tapio Pahikkala, Antti Airola, Harri Merisaari, Jani Saunavaara, Saeid Alinezhad, Riina-Minna Väänänen, Terhi Tallgrén, Janne Verho, Aida Kiviniemi, Otto Ettala, Juha Knaapila, Kari T. Syvänen, Markku Kallajoki, Paula Vainio, Hannu J. Aronen, Kim Pettersson, Peter J. Boström, and Pekka Taimen. Prostate Cancer Risk Stratification in Men with a Clinical Suspicion of Prostate Cancer Using a Unique Biparametric MRI and Expression of 11 Genes in Apparently Benign Tissue: Evaluation Using Machine-Learning Techniques. Journal of Magnetic Resonance Imaging, 2020; 51(5): 1540-1553.

V Ileana Montoya Perez, Ivan Jambor, Tommi Kauko, Janne Verho, Otto Ettala, Ugo Falagario, Harri Merisaari, Aida Kiviniemi, Pekka Taimen, Kari

T. Syvänen, Juha Knaapila, Marjo Seppänen, Antti Rannikko, Jarno Riikonen, Markku Kallajoki, Tuomas Mirtti, Tarja Lamminen, Jani Saunavaara, Tapio Pahikkala, Peter J. Boström and Hannu J. Aronen. Qualitative and Quantitative Reporting of a Unique Biparametric MRI: Towards Biparametric MRI-Based Nomograms for Prediction of Prostate Biopsy Outcome in Men With a Clinical Suspicion of Prostate Cancer (IMPROD and MULTI-IMPROD Trials). Journal of Magnetic Resonance Imaging, 2020; 51(5): 1556-67.

VI Ileana Montoya Perez, Harri Merisaari, Ivan Jambor, Otto Ettala, Pekka Taimen, Juha Knaapila, Henna Kekki, Ferdhos L. Khan, Elise Syrjälä, Aida Steiner, Kari T. Syvänen, Janne Verho, Marjo Seppänen, Antti Rannikko, Jarno Riikonen, Tuomas Mirtti, Tarja Lamminen, Jani Saunavaara, Ugo Falagario, Alberto Martini, Tapio Pahikkala, Kim Pettersson, Peter J. Boström and Hannu J. Aronen. Detection of Prostate Cancer Using Biparametric Prostate MRI, Radiomics, and Kallikreins: A Retrospective Multicenter Study of Men with a Clinical Suspicion of Prostate Cancer. Journal of Magnetic Resonance Imaging, 2022; 55(2); 465-477.

1 Introduction

1.1 Data analysis with a limited amount of data

Although we are living in the era of big data or data flood, in many research areas the amount of available data for a study is still limited. The sample size may range between tenths to hundredths of subjects or observations. This data limitation or scarcity is usually due to time and funding constraints, privacy restrictions, subjects availability, and/or shortcomings in data collection. For example, research studies like clinical trials may only be viable for a small sample size because it might require a complex enrollment process or tests that directly affect the time needed to complete them. Furthermore, studies researching rare diseases have the restriction of a small number of subjects with the condition. However, the sample size should not discourage this type of study.

Studies with a small number of subjects present strengths and limitations [1]. Some of the strengths are related to conducting fast subject enrollment, data curation, records review, and ethical and institutional approval. On the other hand, the limitations of small datasets are associated with pitfalls in the data analysis and result interpretation. In particular, the results could mislead interpretation, produce false positives or false negatives, and overestimate the magnitude of an association. Nevertheless, studies such as hypothesis-generating and feasibility studies commonly use a small sample size to avoid expending too many resources while exploring a new hypothesis.

Non-parametric statistical tests are methods used to analyze data when certain assumptions about the underlying population distribution are not met or when the data may not follow a specific parametric distribution [2]. Furthermore, non-parametric statistical tests are recommended for the analysis of small sample sizes [2]. They are also valuable for examining datasets that exhibit issues such as non-normal distribution of the dependent variable, the presence of outliers, and unequal class distribution, among others. Additionally, there has been an observed increase in the use of non-parametric statistical tests over parametric tests in the medical field, even when a large sample size is available for the study [3].

Machine learning (ML) is a discipline that provides algorithms and techniques for building predictive models from a sample [4]. The resulting model embodies the relationship between the observations in the sample and their associated features. The main objective of the model is to make accurate out-of-sample predictions. For

this reason, the quality and quantity of the available data are relevant to the model prediction performance. In other words, the larger and representative the sample is, the better the model will predict on unseen data. However, as we previously noted, they might be reasons for having a limited amount of data. In this case, methodologies for reliable estimation of model performance are of importance.

Resampling methods provide means for estimating the precision of a sample statistic, performing a significance test, or validating a model performance when the available sample is small. These methods reuse the sample data to infer properties of the population without requiring parametric assumptions. The inference is based on repeated sampling within the same sample, requiring heavy computation. Among the resampling methods, we find permutation tests and cross-validation [5].

A permutation test allows computing the statistical significance of a test statistic. It produces the sampling distribution of the test statistic by computing all possible values of the statistic by going through all possible reordering of the sample observations. A permutation test may be the most powerful test for analyzing a small sample, assuming that the observations are exchangeable under the null hypothesis [6].

Cross-validation (CV) is a resampling technique for estimating the prediction performance of an ML model while maximizing the use of the available data. It consists of splitting the data into disjoint subsets, and then using these subsets for training and testing the ML model in a complementary manner [7; 8]. There are many variations of CV, common ones are hold-out, k-fold, leave-one-out, and leave-pair-out. Choosing the proper CV is problem-dependent; aspects such as the aim of the study, dependencies between sample observations, the structure, and the quality of the data, need to be considered to avoid misleading results.

In combination with cross-validation, a metric is calculated to measure the prediction performance of the ML model. There are many evaluation metrics, and choosing one depends on the type of model and the situation under consideration. Generally, the analyst will select one or more metrics that are popular in their research area. Commonly used metrics for regression models are mean absolute error, mean squared error, and R-squared. In classification models, popular metric choices are accuracy, F-measure, and area under the ROC (receiver operating characteristic) curve. For some metrics, the approach used to compute it under cross-validation can positively/negatively bias the measurement, especially for imbalanced datasets.

ROC curves are useful tools for visualizing the performance of classifiers [9]. The ROC curve plots the classifier's true positive rate (TPR) against the classifier's false positive rate (FPR) over all possible cut-off points, illustrating how the TPR and FPR vary together. Hence, ROC curves allow a visual comparison of performance between classifiers at different cut-off points. In addition, the area under the ROC curve (AUC) quantifies the classifier's accuracy in a single value between 0.0 and 1.0. The AUC value does not depend on the prevalence of a condition or cut-off point, making it a good performance measurement. However, AUC estimated from

cross-validation methods may not reflect the reality if it is not properly derived.

It is then clear that resampling techniques combined with non-parametric statistical tests provide options for analyzing ML models learned from small samples. However, it is recommended that these methods are applied carefully; and results are interpreted with caution. These recommendations and the limitations that small samples present seem to imply a need for studying current and new methods for accurately analyzing small samples.

1.2 Motivation of the research

This thesis focuses on building and evaluating ML models trained on clinical trial data for prostate cancer (PCa) detection and characterization. For many years PCa has been the most commonly diagnosed cancer in men and the second most common cause of cancer-related death. However, in approximately half of the newly diagnosed PCa cases, the patients have a low risk of death from the disease [10]. Accurate identification of PCa risk to stratify patients accordingly is necessary to avoid over-treatment and PCa mortality. The traditional diagnostic protocol for PCa presents challenges; therefore, ML models that predict and characterize PCa could support physicians in accurately establishing the risk in men with a clinical suspicion of PCa. For this reason, a proper assessment of the predictive performance of the ML models is crucial to determine their usefulness and applicability.

The traditional diagnosis of PCa involves prostate-specific antigen (PSA) level and digital rectal examination (DRE) findings to decide the need for random biopsies from the prostate (i.e., systematic biopsies). The probability of PCa increases with the rise of the PSA level; however, PCa can be found at all levels of PSA, and no PSA threshold exists that defines the presence of clinically significant PCa (SPCa) [11]. In addition, it has been shown that DRE has low sensitivity for PCa [12; 13]. Taking these into account and the fact that systematic biopsies only provide limited information about the whole gland pathology [14; 15], SPCa cannot be ruled out entirely based on systematic biopsies findings.

Magnetic resonance imaging (MRI) is a non-invasive method that has shown great potential for detecting and characterizing PCa, making it an ongoing area of research [16; 17; 18; 19; 20; 21]. Similarly, other imaging modalities, biomarkers, and genes have also shown potential for detecting PCa and are under intensive investigation [22; 23; 24]. Furthermore, there is a significant interest in exploring the additional value and interaction that attributes from different sources (e.g., clinical variables, features extracted from MRI, and parameters obtained from genes and blood kallikreins) have on the detection and characterization of PCa. ML models developed using these attributes could improve and automate the diagnosis of PCa, and finding the variables or combinations of variables that yield the highest prediction performance for PCa or SPCa is desirable.

Quality annotations of prostate tumors in MRI require the expertise of experienced radiologists, and the process is very time-consuming. Consequently, obtaining studies with a substantial quantity of high-quality annotated PCa MRI datasets can be challenging. Datasets that include not only annotated prostate MRI but also gene expressions, among other variables, are even more elusive. However, in this work, we had access to datasets from approved studies and registered clinical trials conducted at Turku University Central Hospital (TYKS). The datasets consisted of biomarkers and annotated MRIs from patients with clinical suspicion of PCa. In particular, datasets from a single-center trial (IMPROD; NCT01864135) and a multicenter trial (MULTI-IMPROD; NCT02241122) had a vast number of variables that include clinical variables, genes, blood kallikreins, and features derived from different MRI modalities that makes them suitable for a broader analysis. These trials were carried out between 2013-2017 and involved strong collaboration between departments and hospitals.

The availability of these datasets facilitated the analysis of PCa-related attributes for predicting and stratifying PCa or SPCa. Furthermore, it encouraged us to address the challenge of analyzing and learning from size-constrained datasets while achieving clinically meaningful evaluation outcomes, a matter of interest for both the medical and computer science fields.

1.3 Main objectives and research questions

This work aimed to evaluate the capability of features extracted from prostate MRI alone and in combination with other relevant variables for predicting PCa risk using statistical and machine learning methods suitable for analyzing datasets with size constraints. More precisely, this thesis has three main objectives, which are: to evaluate the prediction power that prostate MRI extracted features have in detecting high-risk PCa using cross-validation techniques, to develop an evaluation method for small samples that allows ROC analysis, and to identify and compare relevant predictors among MRI features, clinical variables, gene expressions and kallikreins for PCa detection and stratification. These objectives are represented by the following research questions:

(**RQ1):** How precise are features extracted from prostate MRI in classifying and stratifying PCa?

(**RQ2):** How to improve ROC analysis derived from cross-validation to evaluate models when the size of the available data is small?

(**RQ3):** How well can linear models that combine variables/features from different sources predict and stratify PCa?

The first question is addressed by publications I and II. These publications evaluate several types of radiomic features extracted from MRI. Publication I focuses on evaluating texture features from diffusion-weighted imaging (DWI) parametric maps for PCa detection using the whole gland image. Publication II evaluates radiomic features extracted not only from DWI but from other MRI modalities for differentiating low-risk (not clinically significant) from high-risk (clinically significant) PCa tumors. The second research question studies AUC and ROC curve estimates obtained from different resampling techniques. This topic is covered in Publication III, where a novel cross-validation method for ROC analysis is proposed and tested using PCa data. The last research question is explored in Publications IV, V, and VI. Where PCa-related variables/features are evaluated alone or combined with MRI features to determine their contribution to differentiate benign/low-risk PCa from high-risk PCa. For a summary of the outcomes arranged by research question and original publications see Table 1.

1.4 Structure of the thesis

This thesis consists of two parts. Part I is formed by Chapters 1-5. Chapter 1 provides an introduction to the subject, motivation, and research questions of this thesis. Then, Chapter 2 gives the background information on prostate cancer which is the domain of the data, and conveys the main objectives of this research. Chapter 3 presents a comprehensive overview of the technical foundation and methods used. The research publications and results are summarized in Chapter 4, followed by the conclusion, discussion, and future work in Chapter 5. Part II presents the six original publications that resulted from this work.

Table 1. Main outcomes arranged by research question (RQ) and original publications.

RQ: Publication	Outcomes
RQ1: I, II	• DWI parametric map features showed good predictive performance for PCa. • Combining radiomics from DWI parametric maps and T2W had excellent results in stratifying PCa tumors into low- and high-risk. • T2 mapping radiomics had the lowest classification performance and added little value. • Features derived from GLCM, Gabor filters, and Zernike moments excelled in stratifying PCa tumors.
RQ2: III	• The LOOCV AUC estimate exhibited a negative bias, making it unreliable for ROC analysis. • LPOCV and the new TLPOCV methods provide nearly unbiased AUC estimates. • TLPOCV enables ROC analysis and is more reliable than LOOCV.
RQ3: IV, V, VI	• The bpMRI Likert score from an experienced radiologist exhibited the best predictive performance for PCa and SPCa. • A linear model using selected clinical variables and mRNA transcripts exhibited high SPCa detection performance. • Selected tumor radiomic features had similar SPCa prediction performance to the bpMRI Likert score. • Combination of clinical variables, kallikreins, and WG radiomics showed promise but were not superior to bpMRI Likert score for PCa and SPCa prediction.

PCa: Prostate cancer; DWI: Diffusion-weighted imaging; T2W: T2-weighted imaging; LOOCV: Leave-one-out cross-validation; LPOCV: Leave-pair-out cross-validation; TLPOCV: Tournament LPOCV; WG: whole gland; bpMRI: bi-parametric MRI; SPCa: clinically significant PCa; GLCM: Gray-level co-occurrence matrix.

2 Prostate cancer domain-specifics

2.1 Prostate gland and its anatomy

The prostate is an accessory gland of the male reproductive system. Its primary function is to produce prostatic fluid, which is essential for male fertility, as it contains several factors that control the ejaculation process and the survival of spermatozoa [25]. It is located in front of the rectum and below the urinary bladder, and its shape resembles a truncated cone. The prostate, as described by McNeal J.E. in [26], has four basic anatomic regions:

1. The *peripheral zone* (PZ), which constitutes over 70% of the glandular prostate.

2. The *central zone* (CZ), which constitutes about 25% of the glandular prostate and contains the ejaculatory ducts.

3. The *transition zone* (TZ), which surrounds the urethra and constitutes approximately 10% of the glandular prostate.

4. The *anterior fibromuscular stroma* is a thick non-glandular layer that shields the anterior surface of the three previous glandular regions.

The prostate gland is a direct target of prevalent benign and malignant diseases, such as benign prostatic hyperplasia (BPH) and prostate cancer [25]. About two-thirds of the diagnosed prostate cancers are located in the PZ, and the rest are located primarily in the TZ, while CZ tumors are rarer [27]. BPH, which is a non-cancerous condition of enlargement of the prostate, typically originates in the TZ [26].

2.2 Incidence and mortality of prostate cancer

Prostate cancer is the second most frequently diagnosed cancer and the fifth leading cause of cancer death among men, with an estimate of almost 1.4 million new cases and 375,000 deaths worldwide in 2020 [28]. In Finland, according to the Finnish Cancer Registry 2020, prostate cancer was the most common cancer diagnosed in men, with 5035 new cases, and the second most common cancer to cause cancer-related deaths with 928 fatalities [29].

2.3 Grading of prostate cancer

PCa is commonly evaluated from biopsy or prostatectomy specimens. Tissue samples from the prostate are examined under a microscope, and cells are graded using the Gleason grading system. This grading system was created by Dr. Donald Gleason in 1966 based exclusively on the patterns found on prostate tumors [30]. The grade was defined as the sum of the two most common grade patterns seen in the specimen and reported as the Gleason score (GS) ranging from 2 to 10, although scores lower than six are rarely assigned [30]. However, since its creation, the grading system has been slightly modified, and in 2014, the International Society of Urological Pathology (ISUP) consensus conference proposed a new grading system based on the original GS. In this new grading system, patterns were arranged in five grade groups (Table 2) to provide a more accurate stratification of the tumor and improve PCa prognosis [31].

Table 2. ISUP Gleason Grade Groups [31].

Gleason Grade Group (GGG)	Gleason Score
1	3+3
2	3+4
3	4+3
4	4+4, 3+5, 5+3
5	4+5, 5+4, 5+5

In this thesis, the Gleason grade group $(GGG)$was used when defining the patient-level ground truth. For our analyses, we dichotomized our sample into two groups: benign/clinically non-significant $(GGG < 2)$ vs. clinically significant $(GGG \geq 2)$. In the tumor-based analyses, the tumors were labeled as low if GS $= 3 + 3$ and high otherwise.

2.4 Screening and Diagnosis of prostate cancer

Prostate-specific antigen (PSA) and digital rectal examination (DRE) are screening methods routinely used to estimate the risk of prostate cancer. PSA screening has been associated with overdiagnosis and overtreatment [32; 33], and DRE alone is not recommended for early detection due to its low sensitivity [13]. Nevertheless, it has been said that using PSA along with DRE increases the chance of early detection of PCa [12].

Usually, PCa is suspected when PSA is elevated and/or with an abnormal DRE, and these indications often lead to a prostatic biopsy. However, according to the European Association of Urology (EAU), other considerations need to be taken before a biopsy. For example, the EAU [32] recommends confirming an elevated PSA using the same laboratory and under standardized conditions before considering a biopsy.

In addition, the EAU guidelines presented the free-to-total PSA ratio and a panel of kallikreins (Prostate Health Index test and 4Kscore) for risk stratification as options to avoid unnecessary biopsies. Yet it stresses that a formal comparison of these tests is needed.

Traditionally, after clinical suspicion of PCa, transrectal ultrasound (TRUS) and systematic biopsies were used to diagnose PCa. The EAU guidelines recommend taking ten- to 12-core biopsies and reporting each biopsy site individually. Furthermore, the guidelines mention that an ISUP GGG should be provided as an overall grade. Additionally, the EAU guidelines also presented the role of multiparametric MRI (mpMRI) in PCa diagnosis. Indicating the potential that mpMRI has in reliably detecting aggressive tumors. EAU suggests that systematic and targeted biopsies combined may also be better for predicting the overall GGG. In the included articles of this thesis, the patient-level ground truth is based on the combination of systematic and targeted biopsy or radical prostatectomy if available.

2.5 MRI in prostate cancer diagnostics

Multiple studies have shown the potential of mpMRI in detecting clinically significant PCa and reducing insignificant PCa findings [16; 34; 35]. As a result, the use of mpMRI for detecting and stratifying clinically significant PCa has increased [36]. Moreover, the EAU guidelines recommend mpMRI as support for initial PCa diagnosis if intermediate or high-risk PCa is suspected [32].

Prostate mpMRI combines information from anatomical and functional MRI sequences. The standard prostate mpMRI consists of anatomical T2-weighted (T2W) imaging, DWI, and dynamic contrast enhancement (DCE) sequences. However, the role of DCE in the detection of prostate cancer is under debate [36]. Therefore, a biparametric MRI (bpMRI) consisting of T2W and DWI sequences has become a more appealing alternative [37].

The prostate imaging reporting and data system (PI-RADS) is a reporting scheme designed to standardize image acquisition techniques and interpretation of prostate mpMRI. PI-RADS v2 is widely accepted and used in practice and research, and its latest update version v2.1 was presented by Turkbey et al. in 2019 [36]. In PI-RADS v2.1, clinically significant PCa is defined as GS $\geq$ 7, and/or tumor volume $\geq$ 0.5 cc, and/or extraprostatic extension [38]. Furthermore, PI-RADS v2.1 uses a 5-point scale to assess the correlation between the findings in the mpMRI and the presence of clinically significant PCa as presented in (Table 3) [38].

Alternative to PI-RADS, a five-point Likert system can be used for scoring prostate MRI. The Likert system is not tied to a structured reporting system; hence it may vary between studies. Its score reflects the radiologist's overall impression of clinically significant PCa suspicion. Furthermore, Likert and PI-RADS, as indicated by Khoo et al. in [39], differ in their scoring description, level of analysis, and implementa-

Table 3. PI-RADS v2.1 Assessment Categories [38].

PI-RADS	
1	Very low (clinically significant cancer is highly unlikely to be present)
2	Low (clinically significant cancer is unlikely to be present)
3	Intermediate (the presence of clinically significant cancer is equivocal)
4	High (clinically significant cancer is likely to be present)
5	Very high (clinically significant cancer is highly likely to be present)

tion. Particularly, Likert combines non-prespecified imaging, biochemical data, and reader experience, while PI-RADS evaluation is on pre-specified imaging features in a defined order. The level of analysis in Likert can be patient or tumor-based, while in PI-RADS is tumor-based only. PI-RADS implementation is for detection only, while Likert can be implemented for detection, active surveillance, and recurrence, among others.

A Likert score (i.e, IMPROD bpMRI Likert score) was included in the analyses involved in this thesis as a qualitative feature, the same as the PI-RADS v2.1 score. This Likert scoring for reporting bpMRI was developed in IMPROD clinical trial [21]. More precisely, the Likert score was based on a combined evaluation of T2W, DWI, and the apparent diffusion coefficient (ADC) map. In this Likert system, the assessment categories for clinically significant PCa are analogous to PI-RADS categories presented in (Table 3).

One of the objectives of this thesis was to evaluate features extracted from prostate MRI for predicting and stratifying PCa. Therefore, in the following subsections, as T2W and DWI sequences are part of the gold standard of prostate MRI, we will provide more details on these two MRI modalities.

2.5.1 Anatomic MRI of the prostate

The prostatic zonal anatomy is well depicted in T2W images. In the T2W image of an older man, the PZ zona has a homogeneous high signal intensity, while a younger man can have diffuse intermediate to low signal intensity [40]. The central gland (i.e., CZ and TZ) has a lower signal intensity if compared to the PZ. In the case of prostate cancer, tumors in the PZ appear as an ellipsoid or circular sub-capsular shape of low signal intensity. While in the central gland, tumors appear as homogeneous low signal intensity without a capsule. As an example, Figure 1(a) presents a T2W axial prostate image with a GS 7 tumor in PZ zone.

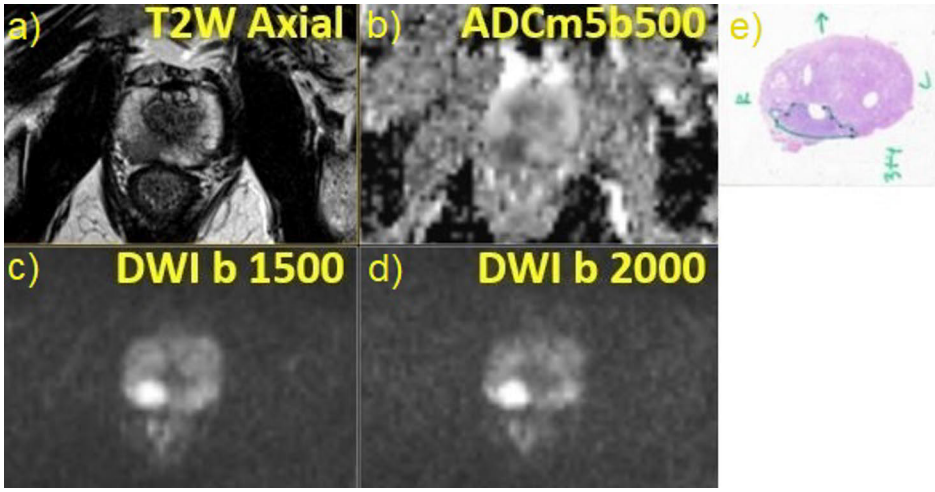

Figure 1. Apperance of the peripheral zone (PZ) prostate cancer in a) T2-weighted imaging (T2W axial), b) Apparent diffusion coefficient map (ADCm5b500), c) Diffusion-weighted imaging b-value of 1500 (DWI b 1500), d) Diffusion-weighted imaging b-value of 2000 (DWI b 2000), and e) Histopathology section indicating Gleason score of 3+4. Figure from [40].

2.5.2 Diffusion-weighted imaging (DWI) of prostate

Most clinical MRI scanners can acquire DWI sequences. In DWI, a so-called b-value determines the strength of diffusion weighting applied. In prostate DWI, each b-value is gathered individually, the resulting images are called trace DWI images and are denoted by their b-value with units s/mm^2 [40]. Examples of prostate trace DWI with b-value 1500 and 2000 are shown in Figure 1(c) and (d), respectively. Several methods for postprocessing trace DWI images exist [20; 41]. One of the most simple and common is to generate an ADC map by fitting the monoexponential function:

$$S(b) = S_0(e^{-bADC_m}), \tag{1}$$

where $S(b)$ is the signal intensity at a particular b-value, S_0 is the signal intensity at $b = 0$, and ADC_m is the diffusion coefficient of the monoexponential function. Other mathematical functions can be used to generate an ADC map from the trace DWI images, for example, the kurtosis function:

$$S(b) = S_0(e^{-bADC_k + \frac{1}{6}b^2 ADC_k^2 K}), \tag{2}$$

where $S(b)$ is the signal intensity at a particular b-value, S_0 is the signal intensity at $b = 0$, ADC_k is the diffusion coefficient of the kurtosis function, and K is the kurtosis term. An example of ADC_m map is presented in Figure 1(b). More about prostate DWI technical details and limitations can be found in [40; 41].

2.6 Biomarkers for prostate cancer

A biomarker is an objective and measurable characteristic of biological processes that may or may not reflect a patient's well-being and clinical condition [42]. Examples of biomarkers range from basic measurements like pulse and blood pressure to more complex laboratory tests. In PCa diagnostics, prostate-specific antigen (PSA) is the most commonly used biomarker. The PSA (a.k.a. hK3) is a member of the tissue kallikrein gene family. It is expressed primarily in the prostate and is the major protein in the seminal fluid, while also found in lower levels in the bloodstream. PSA is produced in the prostate by the secretory epithelial cells and secreted directly to the lumen, where active and inactive PSA is generated [43]. When a portion of the active PSA enters the circulation, it is rapidly bound by protease inhibitors, while other PSA forms circulate as free PSA (fPSA). The secretory epithelial cells are bordered by a layer of basal cells and a basement membrane. A feature of PCa is the disruption of the basal cell layer and the basement membrane, which increases PSA's direct access to the peripheral circulation [43], elevating the amount of PSA in the blood. However, PSA is not a cancer-specific marker, and PSA levels can also be affected by BPH, medications (e.g., 5-alpha reductase inhibitors), advanced age, infection, inflammation, and prostate volume [44; 45]. Therefore, PSA has shown to be a poor marker in PCa diagnosis due to its low specificity and lack of threshold for ruling out the presence of PCa [46]. Consequently, to enhance PSA performance in PCa diagnosis and reduce unnecessary biopsies, biomarkers such as fPSA, free-to-total PSA ratio, PSA density (dPSA), human Kallikrein 2 (hK2), and intact PSA are being investigated [22; 24].

It is of major interest to validate the effectiveness of these biomarkers alone or in combination to improve the detection of high-risk/SPCa and reduce overtreatment. Therefore, this thesis investigates a set of biomarkers for PCa detection and characterization by developing and evaluating ML models using data from single-center and multi-center trials.

2.7 Available datasets

The data used in this doctoral thesis is from an approved study conducted at Turku University Central Hospital (TYKS). The entire study adhered to the guidelines outlined in the Declaration of Helsinki. All the protocols of all sub-studies were approved by the local ethical committee. The approval numbers of the sub-studies are as follows: **PRO3** 80/180/2010, **PRODIF** 112/180/2012, **IMPROD** 113/180/2012 (clinicaltrial.gov, identifier NCT01864135), **MULTI-IMPROD** 180/180/2015 (clinicaltrial.gov, identifier NCT02241122). For the collection of fresh tissue samples from total prostatectomies, the approval numbers are VSSHP ETMK 130/180/2008 and Valvira Dnro 394/05.01.00.06/2009.

In publications I, II, and III, the dataset included subsets of patients from a single-center study PRODIF that had enrolled 72 patients with histologically confirmed PCa by robotic-assisted laparoscopic prostatectomy (RALP). The dataset in Publication IV is based on the single-center clinical trial IMPROD, which enrolled 175 men aged 18 years or older with a clinical suspicion of PCa based on two repeated PSA measurements ranging from 2.5-20.0 ng/mL and/or abnormal DRE. Previous prostate surgery, previous diagnosis of PCa, acute prostatitis, or contraindication for MRI were the exclusion criteria. Publications V is based on the data from IMPROD and the multi-center trial MULTI-IMPROD. In the MULTI-IMPROD trial, 364 men at four different institutions were enrolled using the same criteria as in IMPROD. Publication VI combines data from three clinical trials, PRO3, IMPROD, and MULTI-IMPROD. In Table 4, a summary of the datasets used in this thesis is presented. More details are provided in Chapter 4, where each publication is summarized.

Table 4. Summary of Dataset by Publication.

Publication	N	Features	Ground Truth	Study
I	67	Whole prostate voxel-wise ADC_m, ADC_k and K textures arranged in a grid-like metavoxels	Histologically confirmed PCa by RALP	Single-center
II	62	Textures extracted from manually delineated PCa tumors on DWI (ADC_m, ADC_k, K), T2W, and T2	Histologically confirmed PCa by RALP	Single-center
III	20	Voxel-wise textures from PCa tumors in PZ on ADC_m, ADC_k, and K	Histologically confirmed PCa by RALP	Single-center
IV	80	Clinical variables, mRNA transcripts, and MRI qualitative findings	Systematic and targeted biopsy findings	Single-center
V	499	Clinical variables, and MRI qualitative findings	Systematic and targeted biopsy findings	Multi-center
VI	543	Clinical variables, kallikreins, ADC_m and T2W radiomics, and MRI qualitative findings	Systematic and targeted biopsy, or RALP findings	Multi-center

3 Data analysis

3.1 Statistics and machine learning

In data analysis, statistics and ML provide methods for exploring data to discover useful information. Statistics and ML have many similarities, as they share a common goal which is to extract knowledge from data (i.e., a sample). However, they differ in their purpose [47; 48]. In statistics, data modeling is used for finding significant relationships among the sample variables to infer about the population or explain causation. In contrast, ML models are trained to find patterns and relationships in the sample to make accurate predictions on out-of-sample data. Although statistics is one of the foundations of ML, the relationships learned by an ML model are not usually aimed to explain causality but to uncover generalizable predictive patterns. Nevertheless, statistical inference and ML together can be of value in a research project and point to meaningful conclusions [49].

In statistical modeling, the term model refers to a mathematical representation of a set of assumptions concerning the process that generated the available data or sample and which also applies to other samples coming from the same population. In the case of ML, a model refers to the result obtained when a learning algorithm is applied to the available data. In both cases, the model captures relationships present in the data.

Statistical models that are developed for making inferences could also be used for making predictions. However, their main focus is to understand the mechanisms that generated the data. Therefore, the model evaluation is not focused on prediction performance but on the significance and robustness of the model itself. Furthermore, a set of assumptions are made to apply statistical tests for assessing the model's validity. Hence, the model may have low predictive power as its prediction accuracy on unseen data is not its strength.

In ML, data models are built aiming to make the most accurate out-of-sample predictions, regardless of understanding the underlying mechanisms that generated the data. ML justifies the need of having training and testing datasets to find models with strong prediction capabilities. Therefore, the ML model's prediction makes it possible to determine the best course of action (e.g., treatment assignment). However, the model usually lacks interpretability which makes it difficult to prove relationships within the data.

The separation between statistical inference and ML is under debate, as some methods can be used in both [49]. Nevertheless, in a research project, statistical inference and ML modeling could complement each other [47; 49; 50]. For example, ML models can capture complex patterns and relationships that could improve existing statistical models [48]. They can expose new variables related to the output, which can further be investigated in terms of causality. Moreover, ML models can indicate the level of predictability, as they have higher prediction accuracy than a statistical model built for inferences. In the case of a low predictability level, decisions such as collecting additional data or developing a different approach could be made. In addition, understanding the relevance of the variables and how they affect prediction could provide insight that also supports decisions. Therefore, inferences and predictions are necessary for generating, developing, and testing theories. Combining both leads to more thorough data analysis, where conclusions could be more assertive even in small samples.

3.2 Training a model

In statistics and ML, one can roughly divide the methods for training a model into the classes of parametric and non-parametric. A parametric model learns from the data a fixed set of parameters. These parameters are an approximation of the available data. The number of parameters that the model requires is fixed and known beforehand. The most common parametric model is linear regression. In the case of a non-parametric model, the parameters are adjustable and can change depending on the available data. These models are flexible and with non/fewer assumptions than parametric models. However, they require more data, and their complexity affects interpretability. A common non-parametric model is the K-nearest neighbors.

Depending on the study and structure of the available data, a learning algorithm can be chosen for learning a model. ML provides different learning algorithms, which are commonly categorized into supervised, unsupervised, and semi-supervised learning. These categories are based on the learning method used by the algorithm. In supervised learning, the algorithm requires an output label for each sample unit or observation in the data in order to learn the model, opposite to unsupervised learning where labels are not needed, and to semi-supervised where labels are only available for a small set of the sample.

In this work, due to its nature and the available data, we only employ supervised learning. As indicated previously, in supervised learning, each sample unit has a label or output value; therefore, the sample data would be of the form $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ correspond to the vector of inputs for the ith sample unit and $y_i \in \mathcal{Y} \subset \mathbb{R}$ its output. It is worth highlighting that throughout this thesis, we use the terms features, variables, or independent variables interchangeably to refer to the inputs. Similarly, we may use the terms labels,

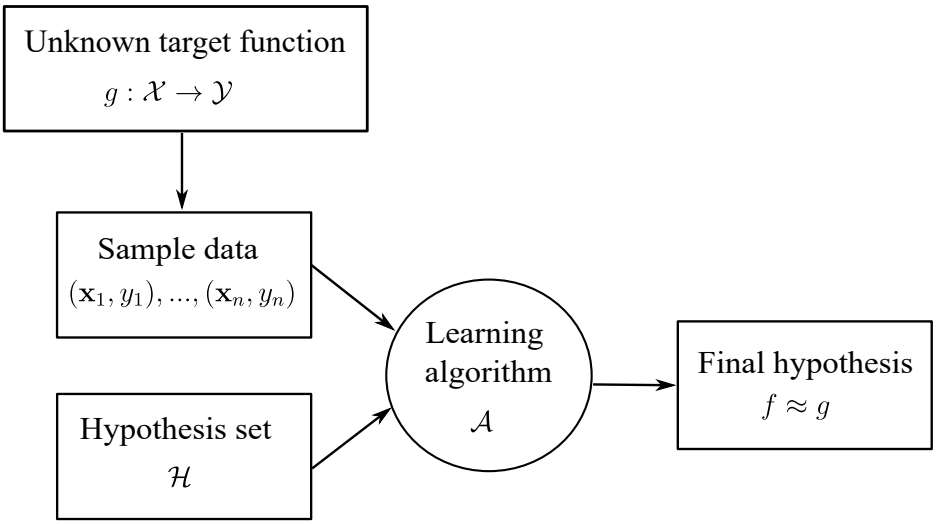

Figure 2. Basic setup of the learning process in ML.

outputs, or dependent variable when referring to the output variable.

In ML, the basic supervised learning setup (Figure 2) as Abu-Mostafa et al., [51] presented it, consists of a learning algorithm $\mathcal{A}$ that uses the dataset $\mathcal{D}$ to choose a function or model f from a set of candidate $\mathcal{H}$ that approximates the unknown target function $g : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. All learning algorithms considered in this thesis find the model f among candidates by minimizing a specific objective function. The objective function, in turn, typically measures how well the candidates perform on the training data, but it may also incorporate some additional measures of hypothesis complexity, in accordance with the idea of simple models generalizing better outside the training set than complex ones.

A popular supervised learning algorithm is **linear regression**. Linear regression performs a regression task that searches for a linear relationship between the independent and dependent variables. The model is parametric and takes the following form:

$$f(\mathbf{x}) = \theta_0 + \sum_{j=1}^{d} x_j \theta_j, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a vector of inputs with d-dimensions, θ_0 is the intercept or *bias* term and θ_j are the model parameters or coefficients.

To find the model $f \approx g$ with optimal parameters $\hat{\boldsymbol{\theta}} = (\hat{\theta_0}, \hat{\theta_1}, ..., \hat{\theta_d})$ the least squares method is commonly used. The objective function which is minimized is the

sum of squared errors (SSE) defined as:

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2 \\ &= \sum_{i=1}^{n} (y_i - \theta_0 - \sum_{j=1}^{d} x_{ij}\theta_j)^2, \end{aligned} \tag{4}$$

where each $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{id})$ is the vector of inputs measured for the ith sample unit, and $\theta = (\theta_0, \theta_1, \theta_2, ..., \theta_d) \in \mathbb{R}^d$ is a vector of parameters. Mathematically, we take the optimal parameters for f to be $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ where $\mathbf{X}$ is the $n \times (d+1)$ matrix with each row being a vector of inputs (a 1 is in the first position for the intercept or *bias* term) and $\mathbf{y}$ the n-vector of real-valued outputs.

Logistic regression is another supervised learning algorithm that learns a linear model, but to solve a classification problem. It is commonly used when the output is formed by two classes (i.e., 1/0, true/false, yes/no, etc.). Like linear regression, the goal is to find the optimal set of parameters $\hat{\boldsymbol{\theta}}$ so that $f \approx g$. However, in logistic regression, there is a functional relationship between the probability of the output class and the inputs. If we take as an example a binary classification with output $y = \{0, 1\}$, the function that describes this relationship is:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-f(\mathbf{x})}} = \frac{e^{f(\mathbf{x})}}{1 + e^{f(\mathbf{x})}} \in [0, 1], \tag{5}$$

where $f(\mathbf{x}) = \theta_0 + \sum_{j=1}^{d} x_j\theta_j$. By definition, all probabilities have to sum up to 1. Hence, the probability of $y = 0$ is $P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}))$.

K-nearest neighbors (KNN) is a non-parametric supervised learning algorithm that can be used to solve regression or classification problems. It consists of making a prediction for a new observation using its K-closest observations in the sample. For regression problems, where the outputs are real values, the mean or median of the K-closest neighbors is used as the predicted value. When KNN is used for classification, the class with the majority of votes from the K-closest neighbors is the predicted class. A distance metric (i.e., Euclidean, Manhattan, etc.) needs to be chosen to determine the proximity of the instances. In addition, the optimal number of neighbors (K) also needs to be defined before using KNN.

As previously stated, when training a model, we aim to find the f that best approximates g. However, the available data $\mathcal{D}$ used in the training process usually contains noise. This noise is random and, in most cases, related to the physical measuring of the data. In the training process, the model may capture irrelevant patterns caused by the noise, misleading and harming the model's inference and prediction performance. Furthermore, there are cases where some or many input variables have

no association with the outputs, including those irrelevant variables in the model adds unnecessary complexity and affects interpretability. Therefore, a set of techniques are available to help with these issues and preserve the generalization capabilities of the model. In the following subsections, feature selection and regularization are presented as examples of these techniques.

3.2.1 Feature selection

In many practical applications, the available data $\mathcal{D}$ may contain redundant, irrelevant, or harmful features. Identifying and removing these features help to solve high dimensionality problems, avoid fitting noisy features, and lower the risk of making a false discovery. Furthermore, selecting a small set of features that have a strong association with the output variable reduces the number of variables that need to be collected and potentially improves prediction accuracy.

In this work, we focus on two categories of feature selection methods: filters and wrappers. Filters methods are generally used, as a preprocessing step, as they select features based on their predictive power or their correlation with the output variable. More precisely, filters evaluate the importance of the features outside a model, providing a ranking of the features based on their individual score obtained with a measure (e.g., mutual information, Pearson's correlation coefficient, AUC). In the case of wrapper methods, an algorithm searches the space of feature subsets with the purpose of finding a set that increases prediction accuracy or reduces the complexity of the model. Classical approaches of wrappers are *backward selection* and *forward selection*. In these approaches, the search algorithm greedily adds or removes features to a model to improve its prediction accuracy. The difference between the approaches is that backward selection starts with all the features, while forward selection starts with no features. Descriptions of these algorithms are available in many books and articles [52; 53; 54; 7].

In the included publications I, II, and VI, feature selection based on filters or wrappers were applied in order to reduce dimensionality by removing irrelevant or noisy features. Additionally, combinations of filters and wrappers were used to improve prediction performance as well as interpretability.

3.2.2 Regularization

Constraining the model training process via penalization limits the model from capturing irrelevant associations or patterns in the data that may have been caused by noise. Further, it helps to avoid overfitting the training data reducing the model variability. As explained earlier, in linear regression, the fitting procedure involves choosing the coefficients that minimize SSE (Equation 4). This process adjusts the coefficients based on the training data, which in the case of a noisy training dataset

results in estimated coefficients that may fail to predict well on new data. Therefore, restraining the learning process by shrinking or regularizing the coefficients toward zero has the effect of reducing variance, and thus increasing the model's generalization capabilities. Two well-known methods for regularizing the regression coefficients are *ridge regression* and *lasso*.

Ridge regression, also known as regularized least-squares (RLS), is a method that minimizes a penalized version of the least squared function [55; 56]. Particularly, the ridge regression coefficients are estimated by minimizing the function:

$$\sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{d} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{d} \theta_j^2 = SSE + \lambda \sum_{j=1}^{d} \theta_j^2, \tag{6}$$

where the term $\lambda \sum_{j=1}^{d} \theta_j^2$, known as the shrinkage penalty, becomes small when $\theta_1, ..., \theta_d$ are close to zero. This penalty has the effect of only allowing the coefficient estimates to become large if there is a proportional reduction in SSE. The parameter $\lambda \geq 0$ is a hyperparameter that determines the amount of penalty to infringe on the coefficient estimates, and it is known as the regularization parameter. If $\lambda = 0$, ridge regression produces the same estimates as least squares, but as $\lambda \to \infty$ the estimates will approach zero. Even though some of the coefficient estimates might become considerably small they will not be set to zero. On the other hand, the least absolute shrinkage and selection operator method known as lasso shrinks some of the coefficients while setting others to zero, thus performing feature selection [57; 8]. The lasso method has a similar formulation as ridge regression. The only difference is that in the lasso the penalty term is replaced by $\lambda \sum_{j=1}^{d} |\theta_j|$, where $\sum_{j=1}^{d} |\theta_j|$ is the l_1 norm. The l_1 norm has the effect of making some estimates to be exactly zero when λ is sufficiently large. Therefore, the lasso method performs feature selection while fitting the model. As a result, the generated model by lasso is easier to interpret than the one produced by ridge regression. As indicated λ is a hyperparameter that can greatly impact the coefficient estimates. For that reason, the value of λ should be carefully selected, for example, by using a resampling technique such as cross-validation, especially in the case of a small sample size.

3.3 Evaluating a model

A methodology for assessing a model's performance is essential for understanding its strengths and weaknesses. For a model to be effective, its results should be measurable, comparable, and reproducible across different samples. Therefore, the performance evaluation must show how well the model generalizes to unseen data.

Classification and regression learning algorithms are highly adaptable, which leads them to learn patterns that are only present in the available data and may not be reproducible. In other words, they can easily overfit the training dataset; hence they

may learn the noise and variability rather than the relationship between variables. It is commonly believed that a model that overfits the training data produces unreliable inferences and predictions. Therefore, in order to determine a model's generalization capabilities, a prediction performance evaluation has to be carried out on an independent test set or by using a resampling technique. In addition, an evaluation metric to indicate the level of prediction accuracy is needed.

3.3.1 Quantitative metrics

Many quantitative metrics exist for evaluating the prediction performance of a model. For example, there are metrics for evaluating performance when the model prediction is numerical and when it is categorical. A commonly used metric for numeric prediction is the mean squared error (MSE). The MSE of a model f is based on the model residuals, which are the deviations between the outputs and the predictions. It is always a positive value that decreases when the predictions approach the true values. More precisely, MSE is computed by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2. \tag{7}$$

In the case of a model predicting a class or the probability of belonging to a class, there are numerous metrics for measuring its performance. For example, if the output corresponds to two classes (i.e., positive class and negative class), a 2×2 contingency table or confusion matrix (Figure 3) is commonly used to evaluate the model. This table or matrix contains the counts of the correct predicted values and the errors made by the model. The four outcomes presented in the confusion matrix are true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Moreover, metrics such as accuracy, sensitivity, specificity, precision, recall, and F1 score, among others, can be derived from those four outcomes. In this work, we mainly focus on ROC curve analysis which results from the TPR $= \frac{TP}{TP+FN}$ and FPR$= \frac{FP}{FP+TN}$ outcomes. Therefore, a more detailed account of ROC curve analysis is given in the following section.

3.3.2 Receiver operative characteristic (ROC) analysis

A ROC curve analysis is based on the trade-off between TPR (i.e., sensitivity) and FPR (i.e., 1- specificity). To compute TPR and FPR, a decision criterion or cut-off point for positivity is needed, making these metrics dependent on the chosen cut-off point. As an alternative, the ROC curve shows the TPR and FPR trade-off over all possible cut-off points (Figure 4). Furthermore, the metrics derived from the ROC curve, such as the AUC, do not depend on the prevalence of a condition or cut-off point [58]. In the ROC space, a curve that represents a perfect classifier is

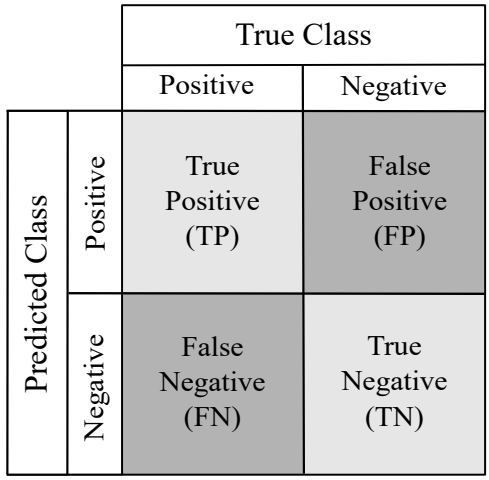

Figure 3. A 2×2 confusion matrix.

the one with a right angle at (0, 1), which means that there is a cut-off point that perfectly separates positives from negatives and the AUC of that classifier is equal to 1. Similarly, a classifier that makes random predictions is represented by a diagonal line from (0, 0) to (1, 1), and the AUC is equal to 0.5. Additional advantages of ROC curves are that several classifiers trained on the same sample can be compared simultaneously at different cut-off points. By visualizing the curve, the sensitivity at a given specificity can be easily obtained.

Regarding the AUC, different approaches for computing it exists [59; 60; 61]. One instance is to estimate the AUC from the ROC curve using the trapezoid rule. An equivalent way is to calculate the Wilcoxon-Mann-Whitney (WMW) statistic by averaging the Heaviside step function scores obtained from all possible comparisons between the pairs of positive-negative data points. The estimated AUC of a function $\hat{A}(f)$ in a finite sample $\mathcal{D}$ using the WMW statistic approach can be formalized as follows:

$$\hat{A}(f) = \frac{1}{|\mathcal{D}_+||\mathcal{D}_-|} \sum_{i \in \mathcal{D}_+} \sum_{j \in \mathcal{D}_-} H\left(f(i) - f(j)\right) \ , \tag{8}$$

where

$$H(a) = \begin{cases} 1, & if \ \ a > 0 \\ 0.5, & if \ \ a = 0 \\ 0, & if \ \ a < 0 \end{cases}$$

is the Heaviside step function, and $\mathcal{D}_+ \subset \mathcal{D}$ and $\mathcal{D}_- \subset \mathcal{D}$ are the positives and negatives, respectively.

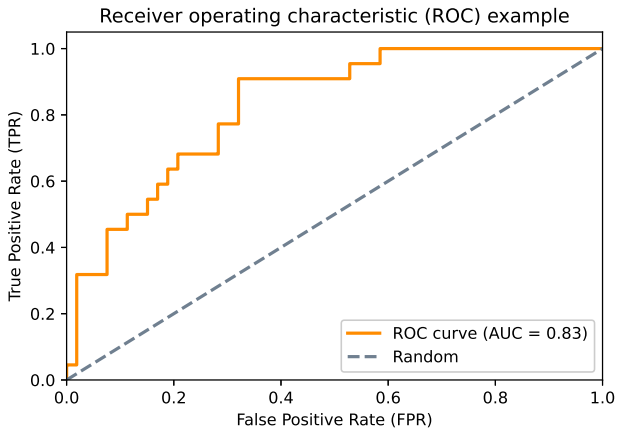

Figure 4. Example of a receiver operating characteristic curve.

3.4 Selecting a model

A set of models can be derived from a sample using different ML algorithms, training with different hyperparameters, or performing feature selection. Some of those models might be better than others for a given task. Therefore, selecting the most appropriate model for reliable and reproducible statistical inference, prediction, or both from a set of candidates is a crucial step in data analysis. Furthermore, the model selection process depends on the objectives or requirements of the study. These requirements might be interpretability, determining feature importance, or producing accurate predictions.

The process of selecting a model involves defining the task that the model is intended to solve, having a set of model candidates for the task, training the models in the available sample data, evaluating the trained models, and selecting the best model according to the evaluation results. Figure 2 can also be interpreted as the model selection process, where the *unknown target function* is the task to be solved, the *sample data* is the available data for training the models, the *hypothesis set* has the model candidates, and the *learning algorithm* is in charge of training and finding the best model for the task, which results in the *final hypothesis* or best model.

Particularly in prediction, the aim of model selection is to find the model with the lowest out-of-sample prediction error from a set of candidates. In order to obtain reliable model performance estimates, an appropriate selection technique matching the aim is necessary. It is well-known that biased estimates emerge if the selection and the evaluation are not performed in separate datasets. Therefore, when the amount of available data is small, techniques based on resampling can aid the selection and evaluation process by maximizing the use of the available data and providing additional information about the fitted models without requiring the model candidates

to be parametric. In the next section, we provide an overview of resampling techniques used in this thesis (i.e., cross-validation and permutation test), and introduce our proposed tournament cross-validation method.

3.5 Resampling techniques

Resampling techniques have become an essential tool in statistics and ML, particularly in analysis where the sample size is small and fail the parametric assumptions [6; 8; 62]. They are methods that involve resampling from a given sample to draw statistical inferences or assess the stability of an estimate. Moreover, these techniques are commonly used for estimating the bias or variance of a statistic or estimator, assessing prediction error, testing statistical hypotheses of an estimated parameter, among other tasks. It is well-known that these techniques are computationally demanding, as they require computing the same statistic or estimator multiple times using different subsets of the sample. However, due to the latest advances in computation power, these techniques are now more feasible in practice.

In this work, we focus on two resampling techniques: cross-validation and permutation test. These resampling approaches do not assume a specific underlying distribution of the observations; thus, they are non-parametric techniques. However, they require that the observations in the sample are independent and identically distributed (IID) or exchangeable. The IID property indicates that the observations are chosen randomly from the same probability distribution. On the other hand, exchangeability indicates that the observations can be rearranged without affecting the underlying probability distribution.

3.5.1 Cross-validation

As explained earlier, proper model selection and performance evaluation are critical for data analysis and machine learning. Therefore, a resampling technique such as cross-validation is crucial to estimate the out-of-sample error associated with an estimator or select the model complexity when the sample size is limited. Cross-validation (CV) involves splitting the sample data and using different splits for training and testing the model when performing a model assessment or complexity selection. Several CV methods exist, and they differ in the number of splits performed on the available data and the aggregation or summarization of the CV results. For example, the hold-out CV is the most basic CV approach. It consists of randomly splitting the data into two sets: training and test sets. In the case of estimating model performance, the training set is used to fit a model, while the test set is used to estimate the efficacy of the model. The K-fold is one of the most commonly used CV. In this CV, the data is divided into K mutually disjoint sets of approximately equal size, which are used as test sets, one at a time, while the remaining K-1 sets are used as

the training set. The model predictions on the K test sets are summarized to estimate the model's performance (e.g., with mean and standard deviation). A variation of the K-fold CV is the popular leave-one-out cross-validation (LOOCV) where K is equal to the sample size (n). Here, each point is held-out at a time as a test set, and the final performance measure is computed from the individual held-out predictions.

The number of splits or folds (K) in a CV is associated with the trade-off of underestimating or overestimating the out-of-sample error of the model trained on the whole dataset. For example, LOOCV uses training sets with $n - 1$ data points, almost as many data points as using the entire available data, making LOOCV an almost unbiased estimator of the model's true performance. In contrast, if a large portion of the available data is set aside as a test set (e.g., hold-out CV) the model's out-of-sample error tends to be overestimated. Here, the training set used to fit the model has much fewer data points than the whole available data, which increases the bias of the estimator. The trade-off between bias and variance associated with the size of K has been studied and explained in [52; 51; 8]. An advantage that splits of larger size (i.e, K $< n$) have is that they are less computationally demanding than, for example, LOOCV. However, the size and the class proportion in the sample may constrain the size of the splits.

Pooling and averaging are two distinct alternatives for aggregating CV results. In pooling, all predictions are grouped as a set, and a performance measure is computed over the predictions set. In contrast, when using averaging, a performance measure is computed for each split in the CV, then all the measurements are averaged as the final result. Both aggregations were presented by Bradley [63] when he studied the use of AUC in evaluating ML algorithms. A noticeable difference between both strategies is that in pooling the predictions from different models are processed together, while in averaging are processed separately which may lead to different AUC.

In K-fold, both aggregation strategies can be used to estimate the AUC. In the pooled K-fold, the AUC is estimated from the set of all the predictions. While in averaged K-fold, an AUC is computed for each test fold, and the final AUC is the average of these fold-wise estimates. In the case of LOOCV, each data point constitutes its test fold, and the AUC is estimated using the pooling approach. Estimating the AUC with the pooling approach is highly risky, as predictions may originate from completely different models, producing biased AUC estimates. Furthermore, several studies performing experiments on simulated and real-world data have shown that both pooled K-fold and LOOCV AUC estimate suffers from high negative bias compared to averaging [64; 65; 66; 67]. For that reason, a CV method that combines the strengths of pooling and averaging, the leave-pair-out cross-validation (LPOCV) for AUC estimation, was proposed [66].

In LPOCV, each pair of positive-negative data points are held-out as a test set, and the CV AUC is computed by averaging over all these pairs' predictions, as in equation (8). This ensures that only pairs from the same CV round are compared,

while it makes maximal use of the available training data. The LPOCV estimate is formally defined as

$$\hat{A}_{LPOCV}(f) = \frac{1}{|\mathcal{D}_+||\mathcal{D}_-|} \sum_{i \in \mathcal{D}_+} \sum_{j \in \mathcal{D}_-} H\left(f_{\mathcal{D}\setminus\{i,j\}}(i) - f_{\mathcal{D}\setminus\{i,j\}}(j)\right) \ ,$$

where $f_{\mathcal{D}\setminus\{i,j\}}$ is the model trained without the i-th and j-th data points.

Although LPOCV produces a more reliable AUC estimate than the pooling version of K-fold and LOOCV, it fails to provide a ranking for each data point prediction, which is needed for a ROC analysis. Therefore, in publication III the tournament leave-pair-out cross-validation (TLPOCV) is proposed as an alternative.

3.5.2 Tournament cross-validation for ROC analysis

A ROC analysis is based on a rank of the data points, where higher ranks are likely to belong to the positive class. In a CV, this is only possible if each data point gets a corresponding model's prediction with a meaningful rank (e.g., probability of belonging to the positive class). As mentioned in the previous section, LPOCV produces an almost unbiased AUC but fails to provide the required ranking for ROC analysis. For that reason, we proposed TLPOCV which is an LPOCV variant that produces a ranking of the model's predictions by combining the method of paired comparisons [68] and round robin tournament theory [69].

In TLPOCV, all possible pairs of data points are held out as test data at a time, including those pairs that belong to the same class. Hence, the number of rounds or paired comparisons carried out is $n(n-1)/2$. Then we considered a tournament graph which is a complete asymmetric directed graph. The tournament graph structure is based on a round-robin competition where participants (i.e., vertices) play each other only once and accumulate points if they win or none otherwise. In the TLPOCV tournament graph, the vertices are the data points, and the edge direction is determined by the predictions' order produced in the train-test split with test set $\{i, j\}$. For example, the direction connecting data points i and j goes from the former to the latter if $f_{\mathcal{D}\setminus\{i,j\}}(i) > f_{\mathcal{D}\setminus\{i,j\}}(j)$. From the graph, we can compute a score for each data point by counting its out-going edges or by the formula:

$$S(i) = \sum_{j=1}^{n} H\left(f_{\mathcal{D}\setminus\{i,j\}}(i) - f_{\mathcal{D}\setminus\{i,j\}}(j)\right) \ .$$

These scores can then be used to estimate the TLPOCV AUC through equation (8). Moreover, by ordering the data points according to their score or number of wins, we obtain the TLPOCV ranking. It has been shown that tournament scores produce a good ranking of the data [70; 71]. Therefore, ROC analysis can be performed with the TLPOCV ranking by using different cutoff points to compute corresponding TPR and FPR.

An issue that might arise in a tournament is inconsistency. This inconsistency emerges in a tournament graph as a cycle. Therefore, it is said that a tournament graph is inconsistent if it has at least one circular triad. In TLPOCV, inconsistency emerges when the learning algorithm is unstable on the sample, such that for data points h, i, and j the following holds:

$$\begin{aligned} f_{\mathcal{D}\setminus\{h,i\}}(h) &< f_{\mathcal{D}\{h,i\}}(i) \\ f_{\mathcal{D}\setminus\{i,j\}}(i) &< f_{\mathcal{D}\setminus\{i,j\}}(j) \\ f_{\mathcal{D}\setminus\{h,j\}}(h) &> f_{\mathcal{D}\setminus\{h,j\}}(j)\ . \end{aligned}$$

From the above situation, we can see that in each of the cases the training data used for learning the function differ by one data point. This difference is enough to produce three functions so different from each other that they create a circular triad. The combination of the available data and the learning algorithm determines how stable is the learning algorithm.

In a tournament graph, the level of inconsistency can be measured by counting the number of circular triads [68; 69; 72]. For example, Kendall and Babington Smith [68] proposed a coefficient of consistency (ζ) for a given complete directed graph, where $0 \leq \zeta \leq 1$. If there are no circular triads in the graph $\zeta = 1$, as the number of circular triads increases ζ tends to zero, and $\zeta = 0$ indicates that the graph has the maximum number of possible circular triads.

If the tournament in TLPOCV is consistent, meaning there are no circular triads in the graph, it generates the strict total order of the data points. Consequently, TLPOCV produces the exact AUC estimate and shares the same unbiasedness property as LPO. However, depending on the severity of TLPOCV tournament inconsistency, both AUCs may drift apart. In publication III, through experiments, we study to what extent this inconsistency affects TLPOCV AUC reliability.

3.5.3 Nested resampling

In addition to model evaluation, CV is also commonly used for model complexity selection. For example, to choose the regularization parameter or to perform feature selection. However, it has been found that using a CV to estimate the out-of-sample error of a model that has been optimized with CV on the same sample significantly biased the estimate [73]. Therefore, to estimate the out-of-sample error correctly, properly defined CV steps should be used when tuning hyperparameters and estimating performance. For instance, the out-of-sample error should be estimated on a large sample independent from the one used in hyperparameter or feature selection. In case of a small sample size, nested resampling, also known as a nested CV, can be used. Nested resampling consists of a series of train/validation/test sets splits, where the train/validation split constitutes an inner loop for tuning parameters and selecting

the optimal model, and an outer loop uses the test set to estimate the selected model out-of-sample error. If the size of the test set in the outer loop is not too large, it gives an almost unbiased estimate of the out-of-sample error.

3.5.4 Permutation tests

To estimate the statistical significance of an observed statistic, a permutation test can be carried out. This test is a resampling technique that obtains the empirical distribution of the observed statistic under the null hypothesis by calculating all possible values of the statistic over the possible rearrangements of the observations in the available sample. As Good P. in [6] presented, a standard permutation test is a five-step procedure:

1. Analyzed the problem—identify the null and the alternative hypothesis.
2. Choose a test statistic that best differentiates the alternative from the null hypothesis.
3. Compute the test statistic for the original sample (e.g., using the original labeling of the observations).
4. Rearrange the observations (e.g., by randomly shuffling the labels), then compute the test statistic again. Repeat until obtaining the distribution of the test statistic for all possible rearrangements or for a large random sample of the rearrangements.
5. Define the significant level for the test and use the obtained distribution in the previous step as a guide to reject or accept the null hypothesis.

For small samples, a permutation test that examines all possible rearrangements is a viable option. However, as the sample size gets larger is more practical and less computationally expensive to utilize the *Monte Carlo method*, which uses the computer to generate a total number of random rearrangements.

Although permutation tests do not require the IID assumption, the observations in the sample must be exchangeable under the null hypothesis. The exchangeability assumption implies that rearranging the observations should not affect their underlying joint distribution. Formally, a sequence of finite random variables $\{X_1, X_2, \ldots, X_N\}$ is exchangeable if the following holds:

$$P(X_1, X_2, \ldots, X_N) = P(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(N)}),$$

where π is any permutation of the indices $\{1, 2, \ldots, N\}$.

A permutation test can be used when assessing the statistical significance of a model or classifier's accuracy [74]. Here, the classifier's test error is the statistic

used to measure the difference between two populations. In every iteration of the permutation procedure, this statistic can be estimated from an independent test set or using cross-validation. In this scenario, the null hypothesis states that we cannot train a classifier that learns the relationship between the variables and the labels, while the alternative is that a classifier can be trained with some accuracy.

In this work, publication IV utilizes a permutation test to assess if the prediction performance of a model trained with only clinical variables could be increased by including an additional set of biomarkers to the model.

4 Research studies and results

4.1 Summary of the publications

In this section, we will go through the six original publications included in Part II of this thesis. For each publication, we provide a summary consisting of the study objectives, motivation, material and methods, results, conclusions, and contribution to the research questions presented in Section 1.3. We also provide the author's contribution to the publication.

4.1.1 Publication I

Diffusion-Weighted Imaging of Prostate Cancer: Prediction of Cancer using Texture Features from Parametric Maps of the Monoexponential and Kurtosis functions.

Objectives: This study introduced a method for the detection of PCa using texture features extracted from DWI parametric maps with a grid approach. The primary objective was to develop an automated system capable of accurately predicting PCa, addressing the challenges associated with prostate DWI evaluation.

Motivation: DWI has demonstrated high diagnostic accuracy for PCa detection. However, quantifying and evaluating prostate DWI can be time-consuming and reader-dependent. This research was motivated by the desire to create an efficient and reliable PCa detection system using texture features extracted from DWI parametric maps.

Materials and Methods: A dataset consisting of 67 patients with histologically confirmed PCa was utilized, with DWI datasets obtained using 12 different b-values ranging from 0 to 2000, and the acquisition time was 8 min 48 seconds. The DWI datasets were fitted using the monoexponential (1) and the kurtosis (2) functions, resulting in three parametric maps: ADC_m, ADC_k, and K. Prostate and tumor delineations were performed using whole-mount prostatectomy sections as ground truth. Texture features, including Gray-Level Co-Occurrence Matrix (GLCM), Local Binary Patterns (LBP), Gabor filter, Haar transform, and Hu moments, were extracted from the three parametric maps. Grid-wise statistical features from ADC_m, ADC_k, and K parametric maps were also calculated. In order to extract the features,

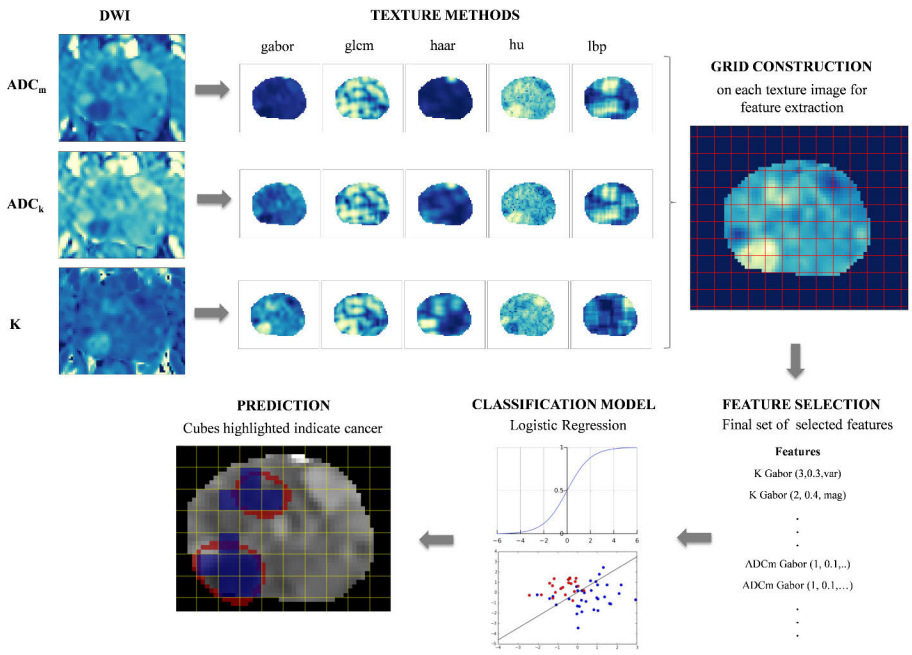

Figure 5. Method pipeline for classifying prostate cubes based on selected texture features extracted from the DWI parametric maps (ADC_m, ADC_k, and K). Figure from publication I [75]. Copyright ©2016, IEEE.

parametric and texture maps were partitioned into equal-sized cuboid regions. The feature was the median value of the voxels inside each cube. The cubes containing 50% or more prostate voxels were considered and marked as cancerous if 10% or more voxels were segmented as part of a tumor or otherwise marked as benign. A total of 12613 prostate cubes and 893 features from each parametric map were obtained to examine the classification performance of features from DWI datasets. Regularized logistic regression was the learning algorithm used to build models for classifying cancerous and benign prostate cubes. The algorithm was applied with two regularization techniques (i.e., l_1 and l_2 norm) to compare their performance. Due to a large number of extracted features and to find a subset of relevant features, two feature selection methods were considered; a filter method based on features' AUCs and a wrapper with a backward selection strategy. In both selection methods, 5% of the total features were selected based on their prediction performance. Leave-subject-out cross-validation (LSOCV) was performed to evaluate the performance of the classifiers. The LSOCV consisted of rounds in which cubes associated with a single patient were held out as test data while the remaining cubes were used for training a model to make predictions on the test data. An AUC was calculated for each patient, and the classifier's overall performance was the average of these AUCs. The feature selection was performed as part of the LSOCV. Figure 5 presents the

method pipeline, where selecting the features and training the model are performed using the training data.

Results: The results of the model evaluation process showed that the highest prediction performance (LSOCV AUC = 0.85) was obtained by backward feature selection when combining the feature sets from ADC_m, ADC_k, and K. Moreover, linear classifiers trained with features extracted from each of the DWI parametric maps using the grid approach resulted in LSOCV AUCs ranging from 0.76 to 0.85, showing the method's potential in differentiating cancerous tissue from benign tissue.

Conclusions: The presented method demonstrates a promising approach to automated PCa detection using DWI parametric maps and advanced feature extraction techniques. The results indicate that this system can effectively differentiate cancerous and benign prostate cubes. Feature selection methods contribute to model simplification, and LSOCV provides a robust assessment of classifier performance on unseen subjects.

Contribution to the research question: This publication contributes to the research question (**RQ1**) by providing quantitative results on the performance of features extracted from DWI parametric maps for predicting PCa.

Author's contribution

Preprocessing and merging the datasets, performing the modeling and evaluation of the classifiers, visualization, and writing the manuscript. The modeling and the evaluation were implemented using self-made Python code and the scikit-learn libraries.

4.1.2 Publication II

Radiomics and Machine Learning of Multisequence Multiparametric Prostate MRI: Towards Improved Non-invasive Prostate Cancer Characterization.

Objectives: The aim of this work was to develop and validate a classification system for predicting PCa Gleason score using radiomic features extracted from three MRI modalities. It explores the combination of imaging modalities and texture extraction methods to determine which ones can effectively assess PCa tumor aggressiveness.

Motivation: PCa is a prevalent and heterogeneous disease, and Gleason score, a common grading system for PCa, is a critical factor for assessing tumor patterns and indicating the level of aggressiveness. Radiomics provides a promising approach to non-invasively assess PCa characteristics, but the optimal combination of MRI

modalities and texture extraction methods for accurate classification remains unclear.

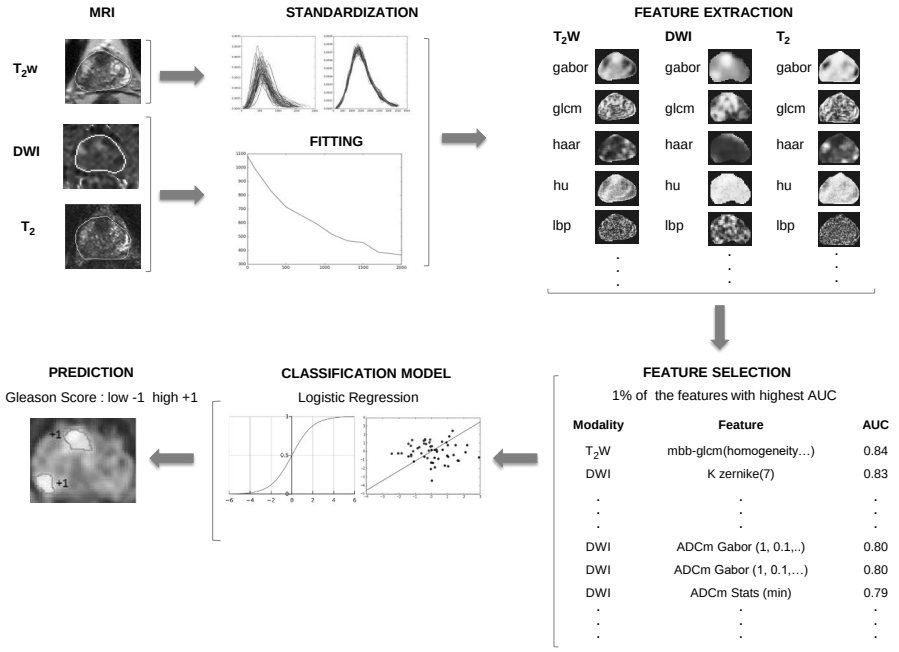

Figure 6. The study pipeline. The T2-weighted images (T_2W) are standardized, the monoexponential and kurtosis functions are fitted to the diffusion weighted images (DWI), and the T2-mapping (T_2) relaxation values are obtained using two parameters monoexponential function. Textures are extracted subsequently. The top 1% of the features are selected by AUC. A regularized logistic regression model is fitted to the selected features and is used to predict the tumor's Gleason score class. Figure and caption from publication II [76].

Materials and Methods: The study utilized MRI datasets comprising of T2W, DWI (12 b-values, 0-2000 s/mm^2), and T2 of 62 patients with histologically confirmed PCa. Tumors were manually delineated on each MRI modality using anatomical landmarks to align the modality to the whole-mount prostatectomy section. The MRI modalities were processed before radiomic feature extraction. T2W was standardized, DWI datasets were fitted with the monoexponential and the kurtosis functions (i.e., ADC_m, ADC_k, and K parametric maps were obtained), and T2 relaxation values were calculated using a monoexponential function. Texture extraction methods, including GLCM, LBP, Gabor filter, Haar wavelet, Sobel, Hu, and Zernike moments, were applied with a "sliding window" algorithm to obtain 2D texture feature maps per slice from the manually delineated PCa tumor. The resulting texture maps were averaged as a tumor-wise median feature. Statistical features were computed for each MRI image type. Various window sizes and parameter combinations yield a total of 1281 features per DWI parametric maps and 1631 per other modalities (i.e.,

T2W and T2). Figure 6 illustrates a comprehensive study pipeline. The final dataset consisted of 100 PCa tumors (i.e., 20 low GS and 80 high GS) and a total of 7015 features when combining all the feature sets.
Logistic regression with either l_1 or l_2 regularization was used to train classifiers for low vs. high GS classification. The predictive performance of the classifiers was estimated by a nested CV strategy, which consisted of an outer LPOCV and an inner 10-fold CV for features and hyperparameter selection. Here, an LPOCV variation was used; instead of negative-positive pairs, every possible pair of data points was used as test data while the remaining data formed the training set for building a model to classify the test data. In each round of the LPOCV, a filter-based feature selection (i.e., selects approx. 1% of all included features based on their AUC) and a regularization parameter selection were performed using the training set. Specifically, the whole training set was used for feature selection and then transformed accordingly to select the optimum regularization parameter value over a set of options using a 10-fold CV. The model for classifying the two data points held out during the LPOCV round was trained with the selected features and the optimal hyperparameter.

Results: Texture methods ranking in the best 1% per modality are presented in Table 5. The highest classification performance, with an LPOCV AUC of 0.88, was achieved by the l_1 regularized logistic regression model with 1% selected features from the union of T2W, ADC_m, and K feature sets. T2 mapping features contributed minimally to the classification performance. ROC curves for the model and the highest AUC statistical and texture features per modality are shown in Figure 7.

Conclusions: This research successfully developed and validated a classification system for characterizing prostate tumors by distinguishing between low and high Gleason scores using radiomic features from multiple MRI modalities. The combination of T2W, ADC_m, and K features, selected through a rigorous feature selection and regularization process, yielded the highest classification performance. This approach holds promise for improving the non-invasive assessment of PCa aggressiveness, potentially aiding in treatment decision-making and patient outcomes.

Table 5. Textures methods ranked in the best 1%. Table from publication II [76].

Image type	Window sizes	Texture extraction method	AUC range
T2W	27	MBB-GLCM, GLCM, Gabor	0.71-0.84
ADC_m	11	Gabor	0.79-0.80
ADC_k	11	Gabor	0.79-0.80
K	7, 9	Zernike, Gabor	0.78-0.83
T2	15, 19, 27, 31, 35	Zernike, Hu, MBB-GLCM, LBP, GLCM, Gabor	0.71-0.75

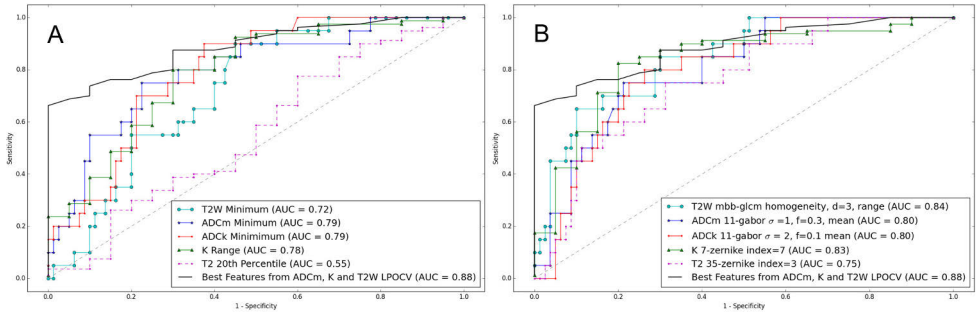

Figure 7. ROC curves within each image type (T2W, ADCm, ADCk, K, T2). A: The best statistical features. B: The best textures features. The final model of the selected features from ADCm, K, and T2W obtained using l_1 regularized logistic regression and validated with leave-pair-out cross-validation is also included in A and B. Figure and caption from publication II [76].

Contribution to the research question: This publication contributes to the research question (**RQ1**) by evaluating the classification performance of statistical and texture features extracted from three MRI modalities for PCa stratification in low- and high-risk tumors. Quantitative prediction estimates of individual features and linear models for classifying PCa tumors were provided as ROC curves and AUC. The ROC curve of the best classifier was obtained from the scores resulting from the LPOCV scores, introducing the method proposed in Publication II which contributes to (**RQ2**).

Author's contribution

Preprocessing and merging the datasets, assessing features individually, performing the modeling and evaluation using nested CV, visualization, and writing the manuscript. The modeling and the evaluation process were implemented using self-made Python code and the scikit-learn libraries.

4.1.3 Publication III

Tournament Leave-Pair-Out Cross-validation for Receiver Operating Characteristic Analysis.

Objectives: This publication introduced a new cross-validation (CV) method called tournament leave-pair-out (TLPOCV). The TLPOCV is a variation of the leave-pair-out cross-validation (LPOCV) that produces a ranking of the data points necessary for ROC analysis. Specifically, TLPOCV constructs a tournament from paired comparisons obtained by carrying out LPOCV overall data point pairs; subsequently, ROC analysis can be performed with the scores determined by the tournament.

Motivation: Other CV methods that allow ROC analysis exist (e.g., LOOCV), but they have been shown to be biased for AUC estimation. The proposed TLPOCV is based on LPOCV, which is currently the most reliable CV method for estimating AUC when the amount of data points is small.

Materials and Methods: Through experiments on synthetic and real-world medical data, AUC estimates of LOOCV, LPOCV, and TLPOCV with two well-established classification methods: ridge regression, and KNN, were empirically evaluated in this publication. In the experiments with synthetic data, the following dataset characteristics: small sample size, class imbalance, low or high dimension, and a large number of irrelevant features, were examined. Furthermore, both signal and non-signal data were considered. Signal data refers to data where the effect size, in this case, the difference between the two distributions from where the classes originate, exists. To produce datasets with signal, the positive and negative classes were sampled from two normal distributions with means one standard deviation apart for at least one of the variables in the dataset. In contrast, for the non-signal data, the data for both classes were sampled from the same normal distribution. In the experiments with real-world data, the dataset consisted of texture features extracted from DWI parametric maps (i.e., ADC_m, ADC_k, and K) of 20 patients with histologically confirmed PCa in the PZ zone. More precisely, the dataset was formed by 85876 voxels (9268 cancerous and 76608 non-cancerous) and six Gabor features that have shown potential in differentiating tumor voxels from non-tumor voxels. In every experiment, the mean and variance of the difference between the CV AUC estimate and the true AUC over several repetitions were computed. The difference was formally defined as $\Delta\hat{A}_{CV}(f) = \hat{A}_{CV}(f) - A(f)$, where CV refers to LOOCV, LPOCV, or TLPOCV, and $A(f)$ is the true AUC. The true AUC in non-signal data is always 0.5, while with signal data the true AUC is not known in advance, but it can be estimated from a large test set drawn from the same distribution that the sample. Therefore, for the experiments with synthetic data, the $A(f)$ was computed from a test set size of 10,000 (5000 positives and 5000 negatives). In the case of real-world data experiments, 30 voxels were sampled to train the model f, and the remaining voxels were used to calculate $A(f)$. In addition to the AUC evaluation, an analogous analysis for estimating sensitivity at a given specificity was carried out to demonstrate the typical case of ROC analysis that is possible by TLPOCV scores.

Results: The results on synthetic data showed that LPOCV and TLPOCV AUC estimates are similar on non-signal and signal data with ridge regression. In the case of KNN, TLPOCV estimates slightly deviate from LPOCV estimates, showing some negative bias. LOOCV compared to LPOCV and TLPOCV presented a larger negative bias in most of the experiment's settings. The variance of the three CV methods

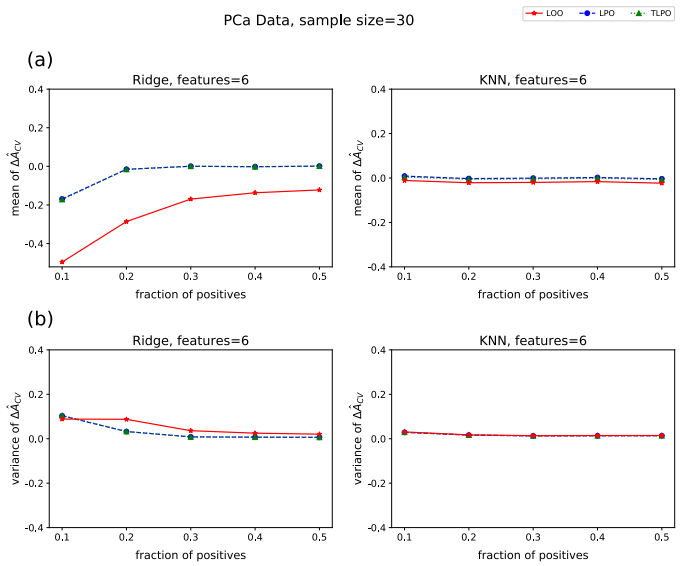

Figure 8. (a) Mean $\Delta\hat{A}_{CV}$ of each cross-validation method on real data as class fraction varies. (b) $\Delta\hat{A}_{CV}$ variances. $\Delta\hat{A}_{CV}$: difference between estimated and true AUC; LOO: leave-one-out; LPO: leave-pair-out; TLPO: tournament leave-pair-out; Ridge: ridge regression; KNN: k-nearest neighbors; PCa: prostate cancer. Figure and caption from publication III [77].

estimates decreased with the increase of the class fraction. Figure 8 shows the result of experiments performed using real-world data. With ridge regression, LPOCV and TLPOCV estimates are almost unbiased, only affected by high-class imbalance, while LOOCV estimates have a strong negative bias. In the case of KNN, LPOCV and TLPOCV estimates are unbiased, and class imbalance seems to not affect the estimates. The variance of all three estimators is close to zero and stable with KNN, while with ridge regression high variance is observed in highly imbalanced class proportions. In the experiments on TLPOCV ROC curve sensitivity at a given specificity, it was observed that the sensitivity tends to be more biased near the ends of the ROC curves (see Figure 9), which is a property of ROC curves calculated from a small sample. For example, the true ROC curve always approaches zero sensitivity for 100% specificity but for ROC curves with a finite sample, this sensitivity may be considerably larger.

Conclusions: TLPOCV, a novel CV method that extends LPOCV, demonstrates its efficacy in AUC estimation, particularly in cases with small data samples. The experiments on synthetic and real-world data showcased its potential to provide reliable AUC estimates. Notably, TLPOCV performs competitively with LPOCV and outperforms LOOCV, which tends to exhibit a more significant negative bias. TLPOCV also demonstrates its applicability in ROC analysis for estimating sensitivity at a

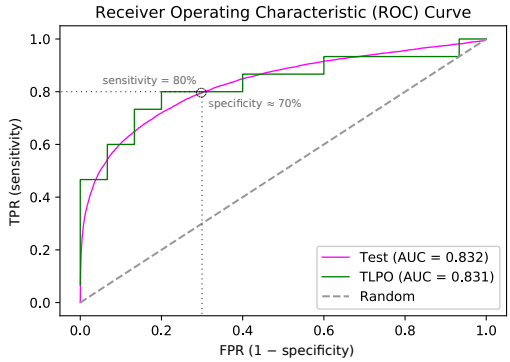

Figure 9. Example of ROC curves of a classifier evaluated by tournament cross-validation (TLPO) and by a large test dataset (Test). The TLPO curve was obtained from 30 random sample units (15 positives and 15 negatives) and the rest of the data was used for the Test curve. The real medical dataset and ridge regression were used. Figure and caption from publication III [77].

given specificity. This study underscores the importance of selecting the appropriate CV method when evaluating classification algorithms, especially in medical and diagnostic contexts.

Contribution to the research question: This publication contributes to the research equation (**RQ2**) by providing a novel cross-validation method that can be used for performing ROC analysis to evaluate the prediction performance of a model trained on a small sample and makes maximum use of the data in the process.

Author's contribution

Developing and executing all the experiments, formal analysis, visualization, and writing the manuscript. The algorithms were developed in Python. Ridge regression was implemented using the RLScore package [78] and KNN using the scikit-learn library.

4.1.4 Publication IV

Prostate Cancer Risk Stratification in Men with a Clinical Suspicion of Prostate Cancer using a Unique Biparametric MRI and Expression of 11 Genes in Apparently Benign Tissue: Evaluation using Machine-Learning Techniques.

Objectives: This study aimed to investigate and evaluate the diagnostic accuracy of parameters from the biparametric MRI protocol (IMPROD bpMRI), both individually and in combination with clinical and molecular markers, for the detection of

significant prostate cancer (SPCa) defined as GS $\geq 3+4$ (i.e., $GGG \geq 2$).

Motivation: Despite the increased use of MRI, accurate risk stratification of men with a clinical suspicion of prostate cancer remains challenging. Therefore, the motivation behind this research was to enhance the accuracy of SPCa detection by examining qualitative and quantitative IMPROD bpMRI parameters, clinical variables, and molecular markers to identify the most effective predictors for SPCa diagnosis.

Materials and Methods: The study comprises data from 80 men suspected of having prostate cancer. This dataset included patients with two repeated measurements of PSA (ranging from 2.5-11.0 ng/ml), measurements of free PSA, acquisition of T2W and DWI images, systematic and targeted prostate biopsies (SB and TB, respectively), and available mRNA data for genetic analysis. The dataset had three variable groups:

- **Clinical variables (eight in total)**: Age, PSA, free-to-total PSA (fPSA), TRUS findings, prostate volume measured by TRUS (TRUS-volume), PSA density based on TRUS (dPSA-TRUS), digital rectal examination (DRE), and 5-alpha-reductase inhibitors (5-ARI).
- **mRNA transcripts (11 in total)**: ACSM1, AMACR, CACNA1D, DLX1, PCA3, PLA2G7, RHOU, SPINK1, SPON2, TMPRSS2-ERG, and TDRD1.
- **IMPROD bpMRI parameters (five in total)**: IMPROD bpMRI Likert score, PI-RADS v2.1 score, DWI-based Gleason grade score (DbGGS), MRI-based prostate volume (MRI-Volume), and PSA density based on MRI (dPSA-MRI).

The study used a combination of SB and TB findings as the ground truth for patients' conditions, where GS $< 3+4$ is considered insignificant/benign prostate cancer, and GS $\geq 3+4$ is considered SPCa. Each variable's potential for SPCa detection was analyzed by computing their AUC and corresponding 95% confidence intervals (CI). A multivariate analysis using ML techniques was conducted to evaluate the prediction performance of combined variables from the same or different groups. Regularized least squares (RLS) with regularization parameter one was used to build regression models from combined variables. To evaluate the model's prediction performance, ROC curves and corresponding AUC were obtained using TLPOCV. Moreover, RLS with the greedy forward feature selection method (GreedyRLS) was used to find a set of variables that produce a model with high prediction performance. The GreedyRLS performance was estimated using a nested cross-validation consisting of a TLPOCV outer loop and an LPOCV inner loop. A permutation test was carried out to determine if adding a variable group to a model trained with another variable group while using GreedyRLS would improve the prediction performance even further. Cohen's kappa statistic (κ) between two readers was calculated for the IMPROD bpMRI Likert score and PI-RADS v2.1 score to assess the interreader agreement,

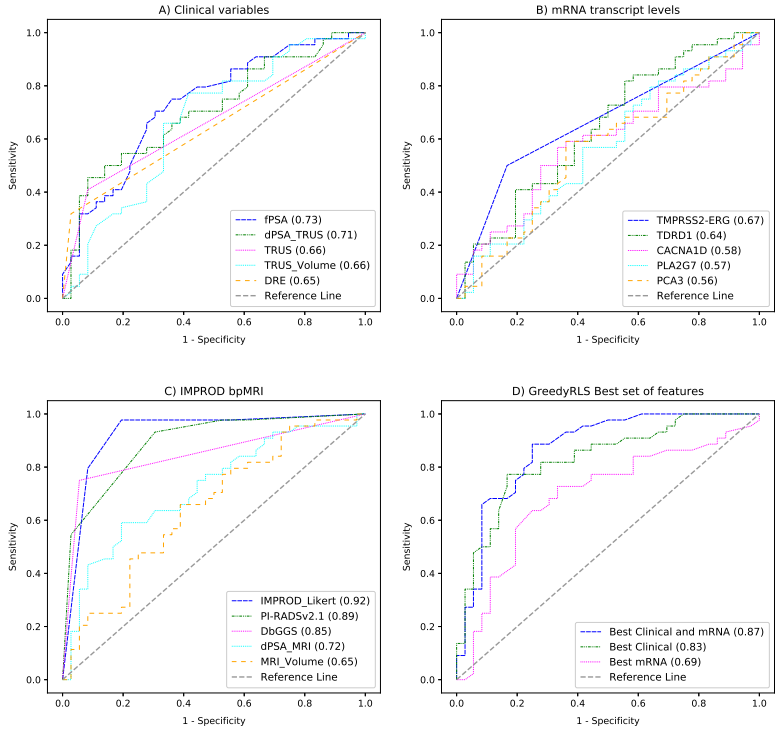

Figure 10. ROC curves for A) clinical variables with highest AUC; B) mRNA transcripts with highest AUC; C) all IMPROD biparametric MRI (bpMRI); D) best set of features obtained using GreedyRLS on the clinical variables, on the mRNA transcripts, and on clinical variables with the mRNA transcripts. Figure and caption from publication IV [79].

and the Spearman rank correlation coefficient (ρ) to evaluate the cross-correlation between the variables.

Results: The study results showed that the IMPROD bpMRI Likert score had the highest prediction performance with TLPOCV AUC = 0.92. The AUC of the clinical variables ranged between 0.56-0.73, while the mRNA transcripts and other IMPROD bpMRI parameters had AUC ranging from 0.50-0.67 and 0.65-0.89, respectively. The feature selection, performed with GreedyRLS using the eight clinical variables and the 11 mRNA transcripts, resulted in the selection of a linear model using fPSA, DRE, TMPRSS2-ERG, PSA, and 5-ARI with a TLPOCV AUC of 0.87, the highest TLPOCV performance without including IMPROD bpMRI Likert score in the model. The permutation test to evaluate if the performance of selected clinical variables could be improved by adding the information of the mRNA transcripts resulted in a p-value of 0.04, which allows rejecting a null hypothesis of no improvement.

Figure 10 presents the ROC curves for the variables with the highest AUC per variable group, and the TLPOCV ROC curve for the best set of features obtained using GreedyRLS on clinical variables, mRNA transcripts, and their combination. The variables that were strongly correlated were dPSA-TRUS and dPSA-MRI ($\rho = 0.91$), TRUS-volumen and MRI-volume ($\rho = 0.88$), and IMPROD bpMRI Likert score and PI-RADS v2.1 score ($\rho = 0.89$). There was a moderate agreement between the two readers in assigning the IMPROD Likert score ($\kappa = 0.57$) and PI-RADS v2.1 score ($\kappa = 0.53$). The agreement improved with the dichotomization of the categories into 1-2 vs 3-5 for IMPROD bpMRI Likert score and for PI-RADS v2.1 score ($\kappa = 0.71$ and $\kappa = 0.65$, respectively).

Conclusions: This study highlights IMPROD bpMRI Likert score as the most effective predictor for significant prostate cancer. The combination of clinical variables and mRNA transcripts in a model demonstrates promising results for SPCa detection. These findings provide valuable insights into enhancing the accuracy of SPCa diagnosis.

Contribution to the research question: This publication contributes to the research equations (**RQ2**) by applying the proposed TLPOCV method in a nested CV for estimating the prediction performance of selected features from different variable groups for predicting SPCa. Contributions are also made to the research question (**RQ3**) as three different sources for PCa biomarkers are evaluated individually and combined.

Author's contribution

Preprocessing and merging the datasets, performing the data analysis, performing the modeling and evaluation using nested CV, applying permutation test for evaluating prediction improvement, visualization, and writing the manuscript. The ML techniques for model evaluation were implemented in Python using the RLScore package [78]. The statistical analyses were conducted using R v. 3.4.3.

4.1.5 Publication V

Qualitative and Quantitative Reporting of a Unique Biparametric MRI: Towards Biparametric MRI-Based Nomograms for Prediction of Prostate Biopsy Outcome in Men with a Clinical Suspicion of Prostate Cancer (IMPROD and MULTI-IMPROD Trials).

Objectives: This study aimed to validate models based on clinical and IMPROD bpMRI data for predicting the presence of prostate cancer (PCa) in prostate biopsy cores. The objectives included developing logistic regression models using different

scenarios and assessing their performance in predicting PCa and SPCa.

Motivation: Accurate prediction of PCa and SPCa is crucial for improving patient outcomes and guiding biopsy strategies. The motivation behind this research was to enhance the accuracy of prediction by incorporating various clinical and IMPROD bpMRI parameters while considering multiple clinical scenarios.

Materials and Methods: The data used in this study was from two registered clinical trials: IMPROD (n=161), and MULTI-IMPROD (n=338). The patient-level ground truth was determined by the findings in targeted and systematic 12-core biopsies. SPCa was defined as biopsy GS $\geq 3 + 4$ ($GGG \geq 2$). Clinical scenarios based on available variables were used to generate logistic regression models. The available variables in each scenario were are follows:

1. **Basic model**: PSA, Age, 5-alpha-reductase inhibitors (5-ARI).
2. **Visit model**: PSA, Age, 5-ARI, DRE.
3. **TRUS model**: PSA, Age, 5-ARI, DRE, TRUS findings, prostate volume based on TRUS (TRUS-volume), PSA density (dPSA-TRUS).
4. **MRI model**: PSA, Age, 5-ARI, DRE, TRUS findings, TRUS-volume, dPSA-TRUS, IMPROD bpMRI Likert or PI-RADS v2.1 score.
5. **MRI model including DWI Gleason grade score**: PSA, Age, 5-ARI, DRE, TRUS findings, TRUS-volume, dPSA-TRUS, IMPROD bpMRI Likert or PI-RADS v2.1 score, DbGGS.

The logistic regression models were developed using the IMPROD dataset and validated on an independent multi-center cohort MULTI-IMPROD. The ability of the models to predict PCa and SPCa was evaluated using AUC with 95% confidence interval. Statistical analysis, such as Cohen's kappa analysis for evaluating the interreader agreement between the MRI group categorical variables (i.e., IMPROD bpMRI Likert score and PI-RADS v2.1 score) and decision curve analysis (DCA) to compare biopsy strategies were performed.

Results: The validation results for models predicting PCa or SPCa in different scenarios are presented in Table 6. Models that included IMPROD bpMRI Likert or PI-RADS v2.1 quality findings had higher prediction performance than the other models. The basic model was the one with the lowest AUC. Cohen's kappa analysis showed a moderate agreement between two readers in assigning IMPROD bpMRI Likert score (κ = 0.59) and PI-RADS v2.1 score (κ = 0.54). When comparing IMPROD bpMRI Likert and PI-RADs v2.1 scoring systems, there was a moderate to substantial agreement in assigning IMPROD bpMRI Likert score and PI-RADS v2.1 score in the IMPROD cohort (κ = 0.63) and MULTI-IMPROD cohort (κ = 0.85). In

the DCA analysis performing a biopsy according to MRI models demonstrated the highest benefit compared to strategies of biopsying no one, according to PSA level, and all men at risk.

Conclusions: This study validated predictive models for PCa and SPCa using clinical and imaging data. The incorporation of IMPROD bpMRI Likert or PI-RADS v2.1 quality findings significantly improved prediction accuracy. Moreover, the study highlights the potential benefit of performing biopsies based on MRI models, which demonstrated the highest utility when compared to alternative biopsy strategies, including those based on PSA levels. These findings contribute to the advancement of prostate cancer diagnosis and biopsy decision-making.

Contribution to the research question: This publication contributes to the research question **RQ3** as it evaluates the performance of models that combined clinical and MRI variables for predicting PCa or SPCa. Here, each model represents a scenario that is determined by the availability of the variables.

Author's contribution

Preprocessing and merging the datasets, performing the data analysis, visualization, and writing the manuscript. The analyses were conducted using R v. 3.5.2.

Table 6. Area Under the ROC Curve (95% Confidence Interval) for five logistic regression models trained on the development cohort (IMPROD trial, N=161) to predict PCa (GS $\geq$ 3+3) or to predict SPCa (GS $\geq$ 3+4) and evaluated on the validation cohort (MULTI-IMPROD, N=338). This Table is part of Table 2 presented in publication V [80].

Logistic Regression model	Benign vs. any PCa	[Benign/3+3] vs. rest of PCa
Basic	0.62 (0.56-0.75)	0.64 (0.58-0.70)
Visit	0.68 (0.62-0.74)	0.75 (0.70-0.80)
TRUS	0.79 (0.74-0.84)	0.80 (0.75-0.85)
MRI: IMPROD bpMRI Likert score	0.86 (0.82-0.90)	0.88 (0.84-0.92)
MRI: PI-RADSv2.1 score	0.87 (0.83-0.91)	0.89 (0.85-0.93)

Basic model = PSA, age, use of 5-alpha-reductase inhibitors.
Visit model = Basic model, and DRE.
TRUS model = Visit model, TRUS findings, prostate volume, and PSA density.
MRI: IMPROD bpMRI Likert score = TRUS model, and IMPROD bpMRI Likert score.
MRI: PI-RADS v2.1 score = TRUS model, and PI-RADS v2.1 score.

4.1.6 Publication VI

Detection of Prostate Cancer using Biparametric Prostate MRI, Radiomics, and Kallikreins: A Retrospective Multicenter Study of Men with a Clinical Suspicion of Prostate Cancer.

Objectives: This study aimed to develop and validate radiomic features derived from biparametric MRI (bpMRI) and kallikreins models for the detection of clinically significant prostate cancer (SPCa, $GGG \geq 2$) using multi-institutional datasets. The study compares the performance of these models with routinely used clinical variables and qualitative IMPROD bpMRI Likert and PI-RADS v2.1 scores.

Motivation: This work was motivated by the need for more accurate and non-invasive methods to detect SPCa. Through the exploration of radiomic features and kallikreins, the aim was to improve the precision of SPCa diagnosis to reduce unnecessary biopsies and the associated risks.

Materials and Methods: The study cohort consisted of 543 men with suspicion of PCa who underwent prostate MRI followed by biopsy as part of a single-center trial or multi-center trial. The ground truth for predicting SPCa was based on biopsy or prostatectomy findings. In Figure 11 the study postprocessing pipeline is presented. Radiomic features were extracted from the manually delineated whole prostate gland (WG) and tumor in ADC maps and T2W images in the initial phase. Next, a pruning and feature selection strategy was applied to obtain a set of radiomic features with high performance in predicting SPCa. Then in the data integration and modeling phase, four variable groups were considered individually and combined. These groups were basic variables (Age, PSA, dPSA, prostate volume), kallikreins (total-PSA, free-PSA, intact-PSA, hK2), MRI qualitative features (IMPROD bpMRI Likert and PI-RADS v2.1 scores), and top selected MRI radiomic features (10 WG and 12 tumor features). In the final phase, variables and models were evaluated on a dataset not used in the previous phases. A univariate analysis consisting of computing AUC (95% confidence interval) was performed to assess the performance of each variable or feature in predicting SPCa. Also, a multivariate analysis using RLS with one as the regularization parameter was conducted to determine the predictive power of combining variables. The analyses were performed in two different data-splitting approaches. In the first approach (split 1), the models were trained using data from a single center (n=72) and externally validated on multi-center data (n=288), whereas, in the second approach (split2), multi-center data were pooled (n=360) and randomly split into 50% for training and 50% for testing. Additionally, models were evaluated using multi-center data (split 1 test data, n=288) and 10-fold CV.

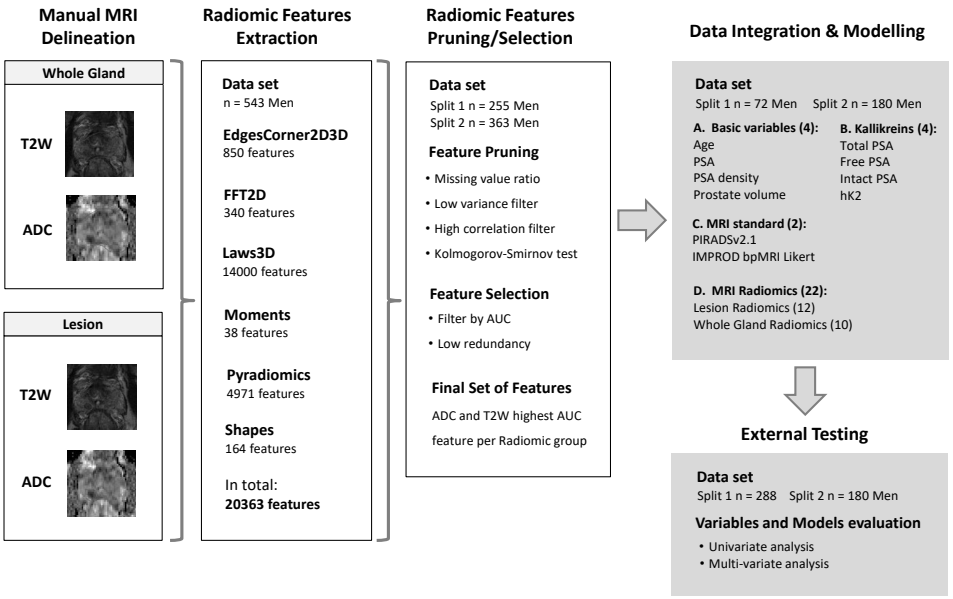

Figure 11. The study pipeline. Figure from publication VI [81].

Results: The evaluation results, regardless of the data splitting approach, showed that the models based on basic variables, kallikreins, and selected WG, alone or combined, had inferior performance in SPCa detection than the qualitative bpMRI score (IMPROD bpMRI Likert score AUC = 0.85) reported by an experienced radiologist. In contrast, selected tumor radiomic features had comparable performance (AUC = 0.83) to IMPROD bpMRI Likert and PI-RADS v2.1 scores. Similar results were observed with 10-fold CV average ROC curves of RLS models for IMPROD bpMRI Likert score, PI-RADS v2.1 score, lesion radiomic features, basic variables, kallikreins, and WG radiomic features using the multi-center evaluation dataset of 288 men of split 1 (Figure 12).

Conclusions: Models based on basic variables, the four kallikreins, and selected WG radiomic features, either individually or in combination, did not outperform the qualitative scores (PI-RADSv2.1 or IMPROD bpMRI Likert) reported by an experienced radiologist. In contrast, a model based on selected lesion radiomic features demonstrated comparable performance to the qualitative scores, with no improvement observed when combined with other variables or features during external validation.

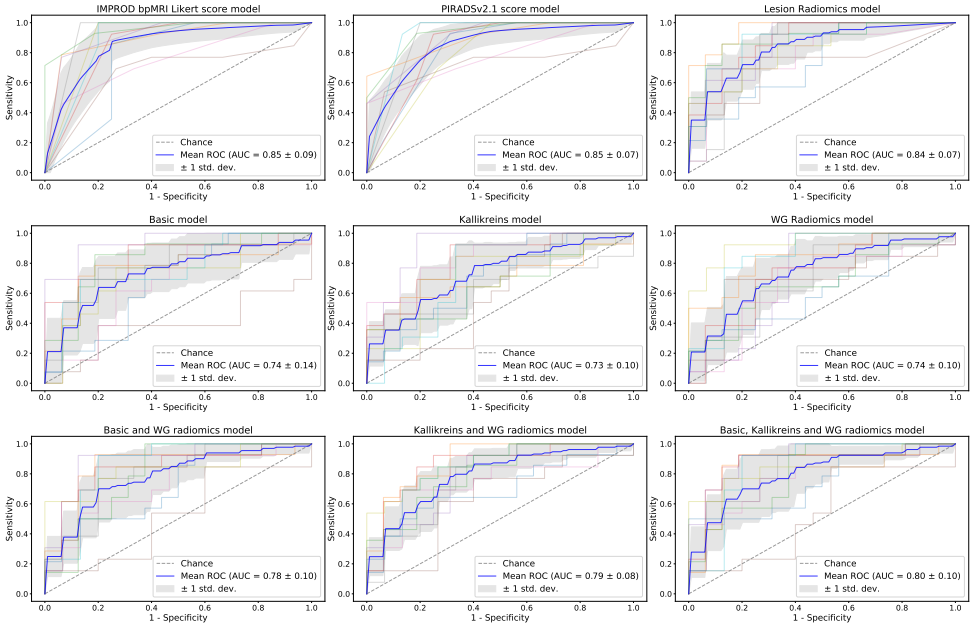

Figure 12. Average 10-fold cross-validation ROC curves for RLS models of IMPROD bpMRI Likert score, PI-RADS v2.1 score, selected lesion radiomics, basic variables, kallikreins, and selected whole gland radiomic features using the data from 288 men in the test set of split 1. Figure and caption from publication VI [81].

Contribution to the research question: This publication contributes to the research question **RQ3** as it evaluates the performance of models that combined clinical, kallikreins, and MRI radiomic features for predicting SPCa. The performance evaluation was performed with a hold-out CV and with a 10-fold CV.

Author's contribution

Preprocessing and merging the datasets, pruning and selecting radiomic features, performing modeling and evaluation, visualization, and writing the manuscript. The modeling and the evaluation were implemented using Python code with RLScore package v 0.8.1 and the scikit-learn library.

4.2 Research results

In this section, each research question is presented and answered according to the research results.

(**RQ1**)**:** How precise are features extracted from prostate MRI in classifying and stratifying PCa?

In publication I, LSOCV results indicate that texture features extracted from DWI parametric maps (ADC_m, ADC_k, and K) have moderate to high performance predicting PCa. Specifically, the highest prediction performance (LSOCV AUC = 0.85) was obtained when selecting features using recursive feature elimination from the three DWI parametric maps. These results confirmed the capability of features extracted from DWI parametric maps in predicting PCa. In the case of stratifying PCa tumors into low-risk and high-risk, LPOCV results in publication II showed that high classification performance can be obtained by a linear model that combines radiomic features from ADC_m, T2W, and K (LPOCV AUC = 0.88). The radiomics extracted for T2 mapping had the lowest classification performance and provided little added value to the linear models. The features based on the GLCM, Gabor filter, and Zernike moments were the most useful for PCa tumor stratification.

(**RQ2):** How to improve ROC analysis derived from cross-validation to evaluate models when the size of the available data is small?

When the available data for training and testing a model is limited, meaning the number of observations is small, cross-validation methods can be used to estimate the model's prediction performance. However, not all cross-validation methods are suitable for ROC analysis. Cross-validation methods that allow ROC analysis, such as LOOCV and pooled K-fold CV, have been proven to be biased for AUC estimation and thus not recommended for ROC analysis. Other cross-validation methods, such as LPOCV, which AUC estimate is almost unbiased, lack the data needed for plotting the ROC curve. Therefore, to improve ROC analysis derived from CV, in publication III, we proposed TLPOCV. A cross-validation method that preserves the advantages of LPOCV for estimating AUC while providing the ranking of the dataset needed for ROC analysis. In the same publication, we empirically evaluated the LOOCV, LPOCV, and TLPOCV estimates on simulated data, where we confirmed the LOOCV AUC estimate bias. Furthermore, through experiments in a real-world dataset of DWI voxels belonging to PCa tumor or non-malignant tissue and six texture features, we demonstrate that LPOCV and TLPOCV AUC estimates are almost unbiased and affected only by a highly imbalanced class distribution in the dataset.

(**RQ3):** How well can linear models that combine variables/features from different sources predict and stratify PCa?

In publication IV, we evaluated the performance of eight clinical variables, eleven mRNA transcripts, and five bpMRI parameters in predicting SPCa. The evaluation result of the individual and combinations of variables showed that the highest performance was by IMPROD bpMRI Likert score assigned by

an experienced radiologist (AUC = 0.92). However, performing feature selection with GreedyRLS when having access to the eight clinical variables and to eleven mRNA transcripts resulted in selecting a linear model form by PSA, fPSA, DRE, TMPRSS2-ERG, and 5-ARI (TLPOCV AUC = 0.87), the highest performance without including IMPROD bpMRI Likert score in the model. In addition, a permutation test to evaluate if a model based on selected variables from the eight clinical could be improved by having access to the eleven mRNA transcripts resulted in a p-value of 0.04, rejecting the null hypothesis of no improvement at a significance level $\alpha = 0.05$.

The combinations of variables for PCa or SPCa prediction were also evaluated in publication V. Here, different scenarios, based on variable availability, were tested on an independent multi-center test data set. The model that resulted in the lowest prediction performance was the basic scenario that consisted of a linear model with age, PSA, and 5-ARI (AUC of 0.62 and 0.64 for PCa and SPCa, respectively). On the other hand, the highest prediction performance resulted when the model included qualitative findings from bpMRI (IMPROD bpMRI Likert or PI-RADS v2.1 scores, AUCs ranging from 0.86 to 0.89).

In publication VI, clinical variables, kallikreins, MRI qualitative variables, selected WG, and tumor radiomic features were evaluated individually and combined for predicting SPCa in men with suspicion of PCa. The evaluation was performed on an independent test set, following two settings: model trained in single-center data and evaluated in multi-center data, and model trained and evaluated in multi-center data. The results in this publication showed, independently from the evaluation settings, that clinical variables, kallikreins, and the selected WG radiomic features, individually or combined, had lower prediction performance than the IMPROD bpMRI Likert score assigned by an experienced radiologist. In contrast, the selected tumor radiomic features have comparable performance to the experienced radiologist. In addition, an average ROC curve obtained by a stratified 10-fold CV on the data set of 288 observations shows the potential that a linear model based on the clinical variables, kallikreins, and the selected WG radiomic features have in predicting SPCa (mean AUC = 0.80).

To summarize, publication IV, V, and VI results showed that a qualitative bpMRI score assigned by an experienced radiologist had the highest prediction performance for PCa or SPCa. However, a linear model based on the selected tumor radiomic features has comparable performance in predicting SPCa. Furthermore, high potential for detecting PCa was observed in a linear model based on selected clinical variables and mRNA transcripts. Meanwhile, a linear model combining clinical variables, Kallikreins, and WG radiomic features showed potential in predicting SPCa.

5 Conclusions

5.1 Summary of the thesis

Chapter 1 covers the thesis introduction, motivation, and research questions. More precisely, it discusses how some research studies have a limited amount of data and the need for caution when performing analysis and inference from small-size datasets. Nevertheless, non-parametric statistics and ML methods that make little assumptions about the data distribution provide tools for analyzing small samples. The proper use and combination of these tools could yield results useful for decision-making, regardless of the sample size. In this thesis, we used the detection and stratification of PCa as our case study for evaluating the performance of variables alone or in combination using statistics and ML methods. The data used in this research is from approved and registered clinical trials conducted in TYKS. The datasets include clinical variables, genes, blood biomarkers, and features derived from different MRI modalities making it suitable for analyzing and developing ML models for PCa detection and stratification. Chapter 2 provides PCa domain-specifics. It presents the function and anatomy of the prostate gland and the incidence and mortality of PCa. The grading of PCa is briefly explained, as this is the base for defining the ground truth or label of the observations needed in the analysis. It also provides the background on screening and diagnosis of PCa, MRI in PCa diagnosis, biomarkers for PCa, and a summary of the datasets. Chapter 3 covers the background of statistical and ML methods used in this thesis. It briefly explains the similarities and differences between statistics and ML methods in data analysis. Highlights the contributions that statistical inference and ML provided to data analysis and how they can complement each other when performing analysis with small sample size. It explains concepts related to deriving a model from data, which include model training, selection, and evaluation. Also, resampling techniques are presented as a viable option for model selection and evaluation when the sample size is small. Furthermore, TLPOCV is proposed as a cross-validation method that allows and improves ROC analysis derived from cross-validation results. In chapter 4, each publication included in this thesis is summarized, and the answers to the research questions are presented. Lastly, this chapter provides the concluding discussion and outcomes of the research.

5.2 Discussion and outcomes

Nowadays, it is still common to find studies with a limited amount of data (i.e., tens to hundreds of observations), as enrollment of subjects might be costly or a complex process, among other reasons. This thesis was motivated by the availability of data from approved clinical trials aiming to improve PCa diagnosis. The data consisted of a vast number of variables (e.g., clinical variables, gene expressions, MRI features, etc.), with the number of subjects ranging from 20 to 543. Therefore, analyzing the capability that these variables have in predicting and stratifying PCa was an objective of this thesis. As we discussed in earlier chapters, non-parametric statistical tests and ML methods provide tools that can be used for analyzing datasets with a scarce number of observations. In our research, in order to accurately evaluate the performance of the available variables in predicting or stratifying PCa, we applied resampling techniques such as CV and permutation test. These techniques allow us to maximize the use of the available data for model selection and evaluation while avoiding strongly biased results. We also propose the TLPOCV, a cross-validation method that allows and improves ROC analysis on CV results. In the following outcomes, we summarize our research findings:

- Radiomic features extracted from DWI parametric maps (i.e., ADC_m, ADC_k, and K) have moderate to high performance in predicting PCa, our analyses confirm the capability that features from DWI parametric maps have in predicting PCa.

- The results for stratifying of PCa tumors into low-risk and high-risk showed that high classification performance can be achieved by combining radiomics extracted from ADC_m, T2W, and K, while features from T2 mapping provided little added value to the classifiers. In addition, the findings suggest that the most useful radiomics for this task were the ones based on GLCM, Gabor filter, and Zernike moments.

- Our results regarding the bias and variance of the LOOCV and LPOCV methods are in line with those presented in earlier work [66; 67], where similar results were also demonstrated for larger sample sizes. We confirmed a substitutional negative bias in LOOCV AUC estimates, which makes it unreliable for ROC analysis. Consequently, we proposed TLPOCV as a more reliable alternative for ROC analysis, provided that the resulting tournament graph remains consistent. A recommended practice is to compute LPOCV and TLPOCV AUC estimates to ensure their similarity before utilizing TLPOCV scores for ROC analysis.

- The qualitative bpMRI score assigned by an experienced radiologist has the highest prediction performance for PCa or SPCa. However, a linear model

based on selected tumor radiomics has a comparable performance for SPCa prediction. Linear combinations of variables from clinical, kallikreins, and MRI WG showed potential for PCa and SPCa prediction but were not superior to the bpMRI score.

The previously stated results were the outcome of properly using CV for model selection and evaluation. In every publication, we selected the CV according to the specified task to reduce bias when evaluating the out-of-sample error of the model trained with the whole dataset. In addition, nested resampling was considered when selecting features to determine their prediction performance, or as an alternative, an independent data set was used. Furthermore, in publication IV, to determine if the improvement of a model trained by greedily combining features from different sources (i.e., clinical variables and mRNA transcripts) was not by chance, we performed a permutation test.

The work presented in this thesis has a few limitations. One of those limitations is that our results are based on a limited amount of available data. Although we applied statistics and ML tools suitable for small sample sizes, validation of our results on an independent large dataset is desirable. In addition, in our studies, we had all MRI datasets evaluated by only one experienced radiologist whose output outperformed all the models that were not based on bpMRI qualitative findings. Thus, it would be of interest to have other radiologists (with different levels of experience) evaluate the MRI datasets to investigate further if models for predicting PCa and SPCa based on clinical, kallikreins, and MRI radiomics can improve or outperform their findings, and determine if these models could provide additional support in detecting PCa or SPCa. Regarding our proposed cross-validation method TLPOCV, one limitation is that it is computationally expensive as the sample size increases, with time complexity of $O\left(n^2\right)$. Therefore, in order to mitigate this limitation, the quicksort leave-pair out cross-validation (QLPOCV) has been presented [82]. The QLPOCV decreases the time complexity to $O\left(n \log n\right)$, making it faster on average than TLPOCV while preserving the advantages of TLPOCV.

Data analysis based on ML models learned from a small amount of available data presents challenges, as it can produce misleading results. In this work, we make substantial contributions to prostate cancer diagnosis and prognosis research while addressing the challenges arising from limited dataset availability. Our research is of significance to data analysts dealing with small datasets and to all medical professionals and healthcare providers engaged in prostate cancer research.

5.3 Future work

Further research on prostate MRI radiomics, clinical variables, genes, and kallikreins is necessary to improve the automatic detection and characterization of SPCa beyond the use of PI-RADS/IMPROD bpMRI as reported by an experienced radiolo-

gist using the qualitative Likert scoring. For example, future work should focus on MRI voxel-wise classification of prostate cancer and automatic prostate segmentation should be integrated into the process. Exploring further variants of the texture features might increase the accuracy of PCa detection and classification. Additionally, it is of interest to compare the regularized logistic regression classifier we employed with other algorithms, such as KNN, random forest, or deep neural networks.

Continuing research on model evaluation methods with a limited amount of data could involve assessing the extent to which TLPOCV results generalize to different data distributions and learning algorithms not covered in this thesis. Furthermore, exploring the application of the TLPOCV rank score to other available ranking metrics, such as Precision and Recall curves, would be of interest.

Finally, in order to support studies with a limited number of subjects or observations, additional research should aim to explore and evaluate methods for analyzing small datasets. This will help ensure that studies with fewer resources can produce reliable results, thus making valuable contributions.

List of References

[1] A. Hackshaw. Small studies: Strengths and Limitations. *European Respiratory Journal*, 32(5): 1141–1143, 2008.

[2] Marjorie A Pett. *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*. Sage Publications, 2015.

[3] Morten W Fagerland. T-tests, Non-parametric Tests, and Large Studies—a Paradox of Statistical Practice? *BMC medical research methodology*, 12(1):1–7, 2012.

[4] Ethem Alpaydin. *Introduction to Machine Learning*. MIT press, 2020.

[5] Michael R Chernick. Resampling Methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3):255–262, 2012.

[6] Phillip I Good. *Resampling Methods*. Springer, 2006.

[7] Max Kuhn, Kjell Johnson, et al. *Applied Predictive Modeling*, volume 26. Springer, 2013.

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[9] Tom Fawcett. An Introduction to ROC Analysis. *Pattern recognition letters*, 27(8), 2006.

[10] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer Statistics, 2021. *CA: a cancer journal for clinicians*, 71(1):7–33, 2021.

[11] Ian M Thompson, Donna Pauler Ankerst, Chen Chi, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Ziding Feng, Howard L Parnes, and Charles A Coltman Jr. Assessing Prostate Cancer Risk: Results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute*, 98(8):529–534, 2006.

[12] William J Catalona, Jerome P Richie, Frederick R Ahmann, M'Liss A Hudson, Peter T Scardino, Robert C Flanigan, Jean B Dekernion, Timothy L Ratliff, Louis R Kavoussi, Bruce L Dalkin, et al. Comparison of Digital Rectal Examination and Serum Prostate Specific Antigen in the Early Detection of Prostate Cancer: Results of a Multicenter Clinical Trial of 6,630 men. *The Journal of urology*, 151(5):1283–1290, 1994.

[13] Leen Naji, Harkanwal Randhawa, Zahra Sohani, Brittany Dennis, Deanna Lautenbach, Owen Kavanagh, Monica Bawor, Laura Banfield, and Jason Profetto. Digital Rectal Examination for Prostate Cancer Screening in Primary Care: a Systematic Review and Meta-analysis. *The Annals of Family Medicine*, 16(2):149–154, 2018.

[14] Klaus Eichler, Susanne Hempel, Jennifer Wilby, Lindsey Myers, Lucas M Bachmann, and Jos Kleijnen. Diagnostic Value of Systematic Biopsy Methods in the Investigation of Prostate Cancer: a Systematic Review. *The Journal of urology*, 175(5):1605–1612, 2006.

[15] Yipeng Hu, Hashim U Ahmed, Tim Carter, Nimalan Arumainayagam, Emilie Lecornet, Winston Barzell, Alex Freeman, Pierre Nevoux, David J Hawkes, Arnauld Villers, et al. A Biopsy Simulation Study to Assess the Accuracy of Several Transrectal Ultrasonography (TRUS)-biopsy Strategies Compared with Template Prostate Mapping Biopsies in Patients who have Undergone Radical Prostatectomy. *BJU international*, 110(6):812–820, 2012.

[16] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic Accuracy of Multi-parametric MRI and TRUS Biopsy in Prostate Cancer (PROMIS): A Paired Validating Confirmatory Study. *The Lancet*, 389(10071):815–822, 2017.

[17] Su-Min Lee, Sidath H. Liyanage, Wahyu Wulaningsih, Konrad Wolfe, Thomas Carr, Choudhry Younis, Mieke Van Hemelrijck, Rick Popert, and Peter Acher. Toward an MRI-based Nomogram for the Prediction of Transperineal Prostate Biopsy Outcome: A Physician and Patient Decision Tool. *Urologic Oncology: Seminars and Original Investigations*, 35(11), 2017.

[18] Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. MRI-targeted or Standard Biopsy for Prostate-cancer Diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

[19] Olivier Rouvière, Ivo G. Schoots, and Nicolas Mottet. Multiparametric Magnetic Resonance Imaging Before Prostate Biopsy: A Chain is Only as Strong as its Weakest Link. *European Urology*, 75(6):889–890, 2019.

[20] Ivan Jambor, Harri Merisaari, Pekka Taimen, Peter Boström, Heikki Minn, Marko Pesola, and Hannu J Aronen. Evaluation of Different Mathematical Models for Diffusion-weighted Imaging of Normal Prostate and Prostate Cancer using High b-values: A Repeatability Study. *Magnetic resonance in medicine*, 73(5):1988–1998, 2015.

[21] Ivan Jambor, Peter J Boström, Pekka Taimen, Kari Syvänen, Esa Kähkönen, Markku Kallajoki, Ileana Montoya Perez, Tommi Kauko, Jaakko Matomäki, Otto Ettala, et al. Novel Biparametric MRI and Targeted Biopsy Improves Risk Stratification in Men with a Clinical Suspicion of Prostate Cancer (IMPROD Trial). *Journal of Magnetic Resonance Imaging*, 46(4):1089–1095, 2017.

[22] Carsten Stephan, Klaus Jung, Michael Lein, and Eleftherios P Diamandis. PSA and other Tissue Kallikreins for Prostate Cancer Detection. *European Journal of Cancer*, 43(13):1918–1926, 2007.

[23] Saradwata Sarkar and Sudipta Das. A Review of Imaging Methods for Prostate Cancer Detection: Supplementary Issue: Image and Video Acquisition and Processing for Clinical Applications. *Biomedical engineering and computational biology*, 7:BECB–S34255, 2016.

[24] Alexander Kretschmer and Derya Tilki. Biomarkers in Prostate Cancer–Current Clinical Utility and Future Perspectives. *Critical reviews in oncology/hematology*, 120:180–193, 2017.

[25] Paolo Verze, Tommaso Cai, and Stefano Lorenzetti. The Role of the Prostate in Male Fertility, Health and Disease. *Nature Reviews Urology*, 13(7):379–386, 2016.

[26] John E McNeal. The Zonal Anatomy of the Prostate. *The prostate*, 2(1):35–49, 1981.

[27] Ronald J Cohen, Beverley A Shannon, Michael Phillips, Rachael E Moorin, Thomas M Wheeler, and Kerryn L Garrett. Central Zone Carcinoma of the Prostate Gland: A Distinct Tumor Type with Poor Prognostic Features. *The Journal of urology*, 179(5):1762–1767, 2008.

[28] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[29] Finnish Cancer Registry. Statistics, Cancer in Finland. URL `https://cancerregistry.fi/statistics/cancer-in-finland/`.

[30] Jonathan I Epstein, William C Allsbrook Jr, Mahul B Amin, Lars L Egevad, ISUP Grading Committee, et al. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *The American journal of surgical pathology*, 29(9):1228–1242, 2005.

[31] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of Prostatic Carcinoma. *The American journal of surgical pathology*, 40 (2):244–252, 2016.

[32] Nicolas Mottet, Joaquim Bellmunt, Michel Bolla, Erik Briers, Marcus G Cumberbatch, Maria De Santis, Nicola Fossati, Tobias Gross, Ann M Henry, Steven Joniau, et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *European urology*, 71(4):618–629, 2017.

[33] Dragan Ilic, Mia Djulbegovic, Jae Hung Jung, Eu Chang Hwang, Qi Zhou, Anne Cleves, Thomas Agoritsas, and Philipp Dahm. Prostate Cancer Screening with Prostate-specific Antigen (PSA) Test: A Systematic Review and Meta-analysis. *bmj*, 362, 2018.

[34] James E Thompson, Daniel Moses, Ron Shnier, Phillip Brenner, Warick Delprado, Lee Ponsky, Marley Pulbrook, Maret Böhm, Anne-Maree Haynes, Andrew Hayen, et al. Multiparametric Magnetic Resonance Imaging Guided Diagnostic Biopsy Detects Significant Prostate Cancer and Could Reduce Unnecessary Biopsies and Over Detection: A Prospective Study. *The Journal of urology*, 192(1):67–74, 2014.

[35] Jurgen J Fütterer, Alberto Briganti, Pieter De Visschere, Mark Emberton, Gianluca Giannarini, Alex Kirkham, Samir S Taneja, Harriet Thoeny, Geert Villeirs, and Arnauld Villers. Can Clinically Significant Prostate Cancer be Detected with Multiparametric Magnetic Resonance Imaging? A Systematic Review of the Literature. *European urology*, 68(6):1045–1053, 2015.

[36] Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, et al. Prostate Imaging Reporting and Data System version 2.1: 2019 Update of Prostate Imaging Reporting and Data System version 2. *European urology*, 76(3):340–351, 2019.

[37] Kristin K Porter, Alex King, Samuel J Galgano, Rachael L Sherrer, Jennifer B Gordetsky, and Soroush Rais-Bahrami. Financial Implications of Biparametric Prostate MRI. *Prostate cancer and prostatic diseases*, 23(1):88–93, 2020.

[38] American College of Radiology et al. Prostate Imaging Reporting and Data System (PI-RADS®), 2019.

[39] Christopher C Khoo, David Eldred-Evans, Max Peters, Mariana Bertoncelli Tanaka, Mohamed Noureldin, Saiful Miah, Taimur Shah, Martin J Connor, Deepika Reddy, Martin Clark, et al. Likert vs PI-RADS v2: A Comparison of Two Radiological Scoring Systems for Detection of Clinically Significant Prostate Cancer. *BJU international*, 125(1):49–55, 2020.

[40] Ivan Jambor, Alberto Martini, Ugo G Falagario, Otto Ettala, Pekka Taimen, Juha Knaapila, Kari T Syvänen, Aida Steiner, Janne Verho, Ileana M Perez, et al. How to Read Biparametric MRI in Men with a Clinical Suspicious of Prostate Cancer: Pictorial Review for Beginners with Public Access to Imaging, Clinical and Histopathological Database. *Acta Radiologica Open*, 10(11): 20584601211060707, 2021.

[41] Roger Bourne and Eleftheria Panagiotaki. Limitations and Prospects for Diffusion-weighted MRI of the Prostate. *Diagnostics*, 6(2):21, 2016.

[42] Kyle Strimbu and Jorge A Tavel. What are Biomarkers? *Current Opinion in HIV and AIDS*, 5(6): 463, 2010.

[43] Steven P Balk, Yoo-Joung Ko, and Glenn J Bubley. Biology of Prostate-specific Antigen. *Journal of clinical oncology*, 21(2):383–391, 2003.

[44] Thomas A Stamey, Mitchell Caldwell, JOHN E McNEAL, Rosalie Nolley, Marci Hemenez, and Joshua Downs. The Prostate Specific Antigen Era in the United States is Over for Prostate Cancer: What Happened in the Last 20 Years? *The Journal of urology*, 172(4 Part 1):1297–1301, 2004.

[45] Reith R Sarkar, J Kellog Parsons, Alex K Bryant, Stephen T Ryan, Andrew K Kader, Rana R McKay, Anthony V D'Amico, Paul L Nguyen, Benjamin J Hulley, John P Einck, et al. Association of Treatment with 5α-reductase Inhibitors with Time to Diagnosis and Mortality in Prostate Cancer. *JAMA internal medicine*, 179(6):812–819, 2019.

[46] Ian M Thompson, Donna Pauler Ankerst, Chen Chi, M Scott Lucia, Phyllis J Goodman, John J Crowley, Howard L Parnes, and Charles A Coltman. Operating Characteristics of Prostate-specific Antigen in Men with an Initial PSA level of 3.0 ng/ml or Lower. *Jama*, 294(1):66–70, 2005.

[47] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[48] Galit Shmueli. To Explain or to Predict? *Statistical science*, 25(3):289–310, 2010.

[49] D Bzdok, N Altman, and M Krzywinski. Points of Significance: Statistics versus Machine Learning. *Nature Methods 2018a*, pages 1–7, 2018.

[50] Max Welling. Are ML and Statistics Complementary? In *IMS-ISBA Meeting on 'Data Science in the Next*, volume 50, 2015.

[51] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from Data*, volume 4. AMLBook New York, NY, USA:, 2012.

[52] Ron Kohavi and George H John. Wrappers for Feature Subset Selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[53] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[54] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag New York, 2 edition, 2009.

[55] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

[56] Ryan Rifkin, Gene Yeo, Tomaso Poggio, et al. Regularized Least-squares Classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

[57] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[58] Karimollah Hajian-Tilaki. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[59] Thomas A Lasko, Jui G Bhagwat, Kelly H Zou, and Lucila Ohno-Machado. The use of Receiver Operating Characteristic Curves in Biomedical Informatics. *Journal of biomedical informatics*, 38(5):404–415, 2005.

[60] Blaise Hanczar, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward R. Dougherty. Small-sample Precision of ROC-related Estimates. *Bioinformatics*, 26(6):822–830, 2010.

[61] James A. Hanley and B. J. McNeil. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1):29–36, 1982.

[62] Persi Diaconis and Bradley Efron. Computer-intensive Methods in Statistics. *Scientific American*, 248(5):116–131, 1983.

[63] Andrew P Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[64] Brian J Parker, Simon Günter, and Justin Bedo. Stratification Bias in Low Signal Microarray Studies. *BMC bioinformatics*, 8(1):1–16, 2007.

[65] George Forman and Martin Scholz. Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57, 2010.

[66] Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. An Experimental Comparison of Cross-validation Techniques for Estimating the Area Under the ROC Curve. *Computational Statistics & Data Analysis*, 55(4):1828–1844, 2011.

[67] Gordon CS Smith, Shaun R Seaman, Angela M Wood, Patrick Royston, and Ian R White. Correcting for Optimistic Prediction in Small Sata Sets. *American journal of epidemiology*, 180(3): 318–324, 2014.

[68] Maurice G Kendall and B Babington Smith. On the Method of Paired Comparisons. *Biometrika*, 31(3/4):324–345, 1940.

[69] Frank Harary and Leo Moser. The Theory of Round Robin Tournaments. *The American Mathematical Monthly*, 73(3):231–246, 1966.

[70] Don Coppersmith, Lisa K Fleischer, and Atri Rurda. Ordering by Weighted Number of Wins Gives a Good Ranking for Weighted Tournaments. *ACM Transactions on Algorithms (TALG)*, 6 (3):1–13, 2010.

[71] Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B Sorkin. Robust Reductions from Ranking to Classification. *Machine learning*, 72(1): 139–153, 2008.

[72] Saul I Gass. Tournaments, Transitivity and Pairwise Comparison Matrices. *Journal of the Operational Research Society*, 49(6):616–624, 1998.

[73] Sudhir Varma and Richard Simon. Bias in Error Estimation when Using Cross-validation for Model Selection. *BMC bioinformatics*, 7(1):1–8, 2006.

[74] Polina Golland and Bruce Fischl. Permutation Tests for Classification: Towards Statistical Significance in Image-based Studies. In *Biennial international conference on information processing in medical imaging*, pages 330–341. Springer, 2003.

[75] Ileana Montoya Perez, Jussi Toivonen, Parisa Movahedi, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J Boström, Aida Kiviniemi, Hannu J Aronen, Tapio Pahikkala, et al. Diffusion Weighted Imaging of Prostate Cancer: Prediction of Cancer Using Texture Features from Parametric Maps of the Monoexponential and Kurtosis Functions. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[76] Jussi Toivonen, Ileana Montoya Perez, Parisa Movahedi, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J Boström, Jonne Pohjankukka, Aida Kiviniemi, Tapio Pahikkala, et al. Radiomics and Machine Learning of Multisequence Multiparametric Prostate MRI: Towards Improved Non-invasive Prostate Cancer Characterization. *PLoS One*, 14(7):e0217702, 2019.

[77] Ileana Montoya Perez, Antti Airola, Peter J Boström, Ivan Jambor, and Tapio Pahikkala. Tournament Leave-pair-out Cross-validation for Receiver Operating Characteristic Analysis. *Statistical Methods in Medical Research*, 28(10-11):2975–2991, 2019.

[78] Tapio Pahikkala and Antti Airola. RLScore: Regularized Least-squares Learners. *The Journal of Machine Learning Research*, 17(1):7803–7807, 2016.

[79] Ileana Montoya Perez, Ivan Jambor, Tapio Pahikkala, Antti Airola, Harri Merisaari, Jani Saunavaara, Saeid Alinezhad, Riina-Minna Väänänen, Terhi Tallgren, Janne Verho, et al. Prostate Cancer Risk Stratification in Men with a Clinical Suspicion of Prostate Cancer using a Unique Biparametric MRI and Expression of 11 Genes in Apparently Benign Tissue: Evaluation using Machine-learning Techniques. *Journal of Magnetic Resonance Imaging*, 51(5):1540–1553, 2020.

[80] Ileana Montoya Perez, Ivan Jambor, Tommi Kauko, Janne Verho, Otto Ettala, Ugo Falagario, Harri Merisaari, Aida Kiviniemi, Pekka Taimen, Kari T Syvänen, et al. Qualitative and Quantitative Reporting of a Unique Biparametric MRI: Towards Biparametric MRI-based Nomograms for Prediction of Prostate Biopsy Outcome in Men with a Clinical Suspicion of Prostate Cancer (IMPROD and MULTI-IMPROD Trials). *Journal of Magnetic Resonance Imaging*, 51(5):1556–1567, 2020.

[81] Ileana Montoya Perez, Harri Merisaari, Ivan Jambor, Otto Ettala, Pekka Taimen, Juha Knaapila, Henna Kekki, Ferdhos L Khan, Elise Syrjälä, Aida Steiner, et al. Detection of Prostate Cancer using Biparametric Prostate MRI, Radiomics, and Kallikreins: a Retrospective Multicenter Study of Men with a Clinical Suspicion of Prostate Cancer. *Journal of Magnetic Resonance Imaging*, 55(2):465–477, 2022.

[82] Riikka Numminen, Ileana Montoya Perez, Ivan Jambor, Tapio Pahikkala, and Antti Airola. Quicksort Leave-pair-out Cross-validation for ROC Curve Analysis. *Computational Statistics*, pages 1–17, 2022.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2024

ISBN 978-951-29-9645-2 (PRINT)
ISBN 978-951-29-9646-9 (PDF)
ISSN 2736-9390 (PRINT)
ISSN 2736-9684(ONLINE)