Published in final edited form as:
Magn Reson Med. 2020 June ; 83(6): 2293–2309. doi:10.1002/mrm.28058.

Repeatability of radiomics and machine learning for Diffusion Weighted Imaging: Short-term repeatability study of 112 patients with prostate cancer

Harri Merisaari[1,2,3], **Pekka Taimen**[4], **Rakesh Shiradkar**[3], **Otto Ettala**[5], **Marko Pesola**[1], **Jani Saunavaara**[6], **Peter J. Boström**[3], **Anant Madabhushi**[3], **Hannu J. Aronen**[1,6], **Ivan Jambor**[1,7]

[1]Department of Diagnostic Radiology, University of Turku, Turku, Finland [2]Department of Future Technologies, University of Turku, Turku, Finland [3]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio [4]Institute of Biomedicine, University of Turku and Department of Pathology, Turku University, Hospital, Turku, Finland [5]Department of Urology, University of Turku and Turku University hospital, Turku, Finland [6]Medical Imaging Centre of Southwest Finland, Turku University Hospital, Turku, Finland [7]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, USA

Abstract

Purpose: To evaluate repeatability of prostate DWI derived radiomics and machine leaning methods for prostate cancer (PCa) characterization.

Methods: 112 patients with diagnosed PCa underwent two prostate MRI examinations (Scan1 and Scan2) performed on the same day. DWI was performed using 12 b-values (0–2000 s/mm^2), post-processed using kurtosis function and PCa areas were annotated using whole mount prostatectomy sections. 1694 radiomic features including Sobel, Kirch, Gradient, Zernike Moments, Gabor, Haralick, CoLIAGe, Haar wavelet coefficients, 3D analogue to Laws features, 2D contours and corner detectors, were calculated. Radiomics and four feature pruning methods (AUC, Maximum Relevance Minimum Redundancy (MRMR), Spearman ρ, Wilcoxon rank-sum) were evaluated in terms of Scan1-Scan2 repeatability using intraclass correlation coefficient, ICC(3,1). Classification performance for clinically significant and insignificant PCa with Gleason Grade Groups 1 vs >1 was evaluated by AUC in unseen random 30% data split.

Results: The ICC(3,1) values for conventional radiomics and feature pruning methods were in the range of 0.28 to 0.90. The machine learning classifications varied between Scan1 and Scan2 with % of same class labels between Scan1 and Scan2 in the range of 61%−81%. Surface to volume ratio and corner detector based features were among the most represented features with high repeatability, ICC(3,1)>0.75, consistently high ranking using all four feature pruning methods, and classification performance with AUC>0.70.

Corresponding author: Harri Merisaari, PhD, Department of Diagnostic Radiology, University of Turku, Kiinamyllynkatu 4-8, P.O. Box 52, FI-20521 Turku, Finland, Tel:+358 2 313 3896, Fax:+358 2 313 2950, harri.merisaari@utu.fi.

Conclusion: Surface to volume ratio and corner detectors for prostate DWI led to good classification of unseen data and performed similarly in Scan1 and Scan2 in contrast to multiple conventional radiomic features.

Keywords

diffusion weighted imaging; repeatability; radiomics; prostate cancer; feature extraction; shape; corner detection; machine learning; intraclass correlation coefficient; Gleason score

Introduction

Qualitative evaluation by means of visual inspection and interpretation of the medical images is a routine clinical practice, while image-derived measurements have been shown to be promising aides to a radiologist in lesion detection and characterization (1). During the last decade, radiomics (2–4) including textures and machine learning (ML) have been applied extensively in medical imaging in general (5–7). Adoption of these methods in routine clinical practice has been limited by concerns related to poor repeatability and reproducibility (8) due to different degree of noise in the imaging data sets obtained from different imaging sessions (9). An increasing number of research groups are applying radiomics for prostate cancer (PCa) imaging (10–18).

Prostate cancer is the most common solid cancer among men in the western world (19). Development of methods for accurate patient-tailored diagnostic process and treatment planning could have a major impact on improved PCa detection, characterization and treatment planning, ultimately leading to improved outcomes of men with suspected or diagnosed PCa. Diffusion weighted imaging (DWI) is a cornerstone of prostate magnetic resonance imaging (MRI) (20), which has an increasingly important role in PCa detection and characterization (21). Aggressiveness of PCa is histologically determined by Gleason Score which is classified into Gleason Grade Group (GGG) (22). Various radiomics and machine learning methods (ML) have already been developed for PCa detection and characterization e. g. (23,24). However, radiomics and ML methods have not been evaluated in terms of their repeatability, which contributes to reservations of making imaging applications that would use prostate MRI more extensively coupled with machine learning in routine clinical practice. High short-term repeatability is a prerequisite towards a quantitative tailored treatment planning and therapy monitoring, a major determining factor of the potential role of ML for prostate MRI in routine clinical practice in the future (25–28).

Repeatability of prostate DWI has been studied with measurements of raw signal (28–34) using intra-class correlation coefficient ICC(3,1) (35). In addition, repeatability of prostate DWI has already been assessed for advanced DWI models such as VERDICT model, which was evaluated (32) in 14 PCa patients who underwent two repeated prostate MRI examinations. Fedorov et al. (36) calculated ICC values for apparent diffusion coefficient maps calculated using mono-exponential function for tumor region in 15 PCa patients scanned with two different scanners. While differences in study subjects and applied MRI acquisition protocols of these studies make a direct comparison of ICC values difficult (37), these studies demonstrated that raw imaging signal and texture features can be repeatable

(27). Recently, public access for prostate MRI of 15 subjects has been provided by Fedorov et al. (38). However, radiomic features for non-Gaussian DWI models (39–41) have not been evaluated in terms of short-term repeatability using the same scanner parameters.

While the area under the receiver operator characteristic curve (AUC) is conventionally used to evaluate binary classification performance for radiomic features, other techniques have been applied as well, particularly in conjunction with machine learning algorithms for PCa classification. In a study by Khalvati et al. (15), maximum relevance minimum redundancy (MRMR) (42) with classification sensitivity and specificity was used to maximize (AUC) in separation of prostate cancer from three non-cancer classes in 20 subjects, and MRMR was used together with Wilcoxon rank-sum test in (43). Spearman ρ value has already been used in feature selection process (44) as well. To the best of our knowledge, repeatability for feature selection and repeatability of machine learning has not been evaluated for prostate DWI.

In this study, we aim to answer three questions 1. How repeatable are DWI prostate radiomics? 2. Which features are to be considered to be both repeatable and with good performance? 3. How does repeatability affect machine learning performance in developing applications for PCa characterization? We evaluated radiomic features and feature selection methods, and trained classifier with machine learning using prostate DWI data, post-processed using kurtosis function. One-hundred-twelve patients with PCa underwent two MRI examinations performed on the same day. We evaluated the short-term repeatability of radiomics, feature selection methods and ML classifiers based on selected radiomics, for repeatability and PCa characterization.

Methods

The study was approved by the institutional review board and each patient gave written inform consent before enrollment to the study. Between March 2013 and February 2016, 115 patients with histologically confirmed PCa scheduled for robotic-assisted laparoscopic prostatectomy underwent MR examination performed using a 3T MR scanner (Ingenuity PET/MR, Philips, Cleveland, USA) and their whole mount prostatectomy sections were available for scientific studies. In one PCa patient, gonadotropin-releasing hormone antagonist (Degarelix, Ferring Pharmaceuticals) was started 10 days before the MR examination while none of the remaining patients had any hormonal, surgical and/or radiotherapy treatment related to prostate before or at the time of imaging.

Three patients were excluded from the final data set due to the presence of severe motion (n=1) and/or susceptibility artifacts (n=2). Patients' characteristics of the 112 included patients are summarized in Supporting Information Table S1. 51 (45%) of the 112 included patients were part of previous studies optimizing b-value distributions for prostate DWI (29) and evaluating different mathematical models for PCa DWI (45–47).

MR examination

Two prostate MRI examinations (3 Tesla Philips Ingenuity PET/MR, Best, Netherlands) were performed on the same day shortly after each other. Following the first MRI examination, the patient was taken out of the MRI room and asked to rest for 10–15 minutes. Following re-positioning of the patient on the MR table, the second MR examination was performed.

Diffusion weighted imaging was performed with spin echo sequence, single-shot echo-planar read-out, monopolar diffusion gradient scheme, and the following parameters: repetition time/echo time (TR/TE) 3141 ms/51 ms, field of view (FOV) 250×250 mm^2, acquisition matrix 100×99, reconstruction matrix 224×224, slice thickness 5 mm, three diffusion directions per each b-value, diffusion gradient timing (Δ) 24.5 ms, diffusion gradient duration (δ) 12.6 ms, diffusion time ($\Delta - \delta/3$) 20.3 ms, 12 b-values of (number of signal averages) 0 (2), 100 (2), 300 (2), 500 (2) (2), 700 (2), 900 (2), 1100 (2), 1300 (2), 1500 (2), 1700 (3), 1900 (4), 2000 (4) s/mm^2. Transversal T$_2$-weighted images were acquired using a single-shot turbo spin-echo sequence with the following parameters: TR/TE 4668/130 ms, FOV 250×320 mm^2, acquisition matrix size 250×320, reconstruction matrix size 512×672, and slice thickness 2.5 mm. Additional MRI data were collected but not evaluated in the current study (48,49) since test-retest data for the additional MRI sequences were obtained only in a small sub-group of patients.

Histopathological analysis and cancer delineation on MRI

Following robotic-assisted laparoscopic prostatectomy, the prostate gland was fixed in formalin. The whole mount prostatectomy sections were processed at 5–6 mm intervals transversely in a plane perpendicular to the long axis of the prostate gland in a superior-inferior direction, similar to axial images of MRI (47,50). Four μm thick whole-mount sections from each macro-block were cut and stained with hematoxylin and eosin. Each individual tumor focus was graded separately based on International Society of Urological Pathology (ISUP) guidelines (22): Gleason Grade Group (GGG) 1 - Gleason score ≤ 3+3, GGG 2 - Gleason score 3+4, GGG 3 - Gleason score 4+3, GGG 4 - Gleason score 4+4/3+5/5+3, GGG 5 - Gleason score 4+5/5+4/5+5. Any Gleason pattern representing less than 5% of tumor volume was graded as tertiary grade (22). Only tumor foci >0.5 cm in diameter, defined using whole mount prostatectomy sections, were included in the analysis. The hematoxylin-eosin stained histological slides were first reviewed by one board certified staff pathologist and later re-reviewed by one experienced genitourinary pathologist. In case of disagreement between the two genitourinary pathologists, the opinion of a third genitourinary pathologist was asked for and consensus opinion was employed as the "ground truth" determination for diagnosis and grade.

Prostate cancer extent on each MRI acquisition (T2w, DWI) was manually delineated by one research fellow (8 years of prostate MRI experience) working in consensus with the genitourinary pathologist (8 years of prostate pathology experience), using whole mount prostatectomy sections as "ground truth". Anatomical landmarks were used to align each MRI acquisition with mount prostatectomy sections.

Data analyses and modeling

Diffusion weighted imaging data sets were fitted using kurtosis function (34,47,51–53) to address potential non-Gaussian behavior when using high b-values in voxel-by-voxel basis, and apparent diffusion coefficient (ADCk) and kurtosis (K) maps were created. Broyden-Fletcher-Goldfarb-Shanno algorithm (54) was used to fit the kurtosis function to the DWI decay curve. Multiple initializations were used to minimize the possibility of local minima (29), and root mean squared error was used as the goodness of fit measure.

We applied a number of conventionally used feature extraction techniques (12,55) found in the literature such as Sobel (56), Kirch (57), Gradient magnitude, Zernike Moments (58), Gabor (59), Haralick (60), CoLlAGe (61), Haar wavelet coefficients (62) and Laws features (63) as texture methods. Additionally, features calculated with edge detector algorithms (64,65) and features derived from corner detectors were obtained.

First-order statistics were calculated (mean, standard deviation, median, range, 25th and 75th percentile) from ADCk and K signal intensities and included as part of feature extractions. Features relating to the shape of the lesion in 2D were acquired as curvature of level-set regions applied on transaxial slices of the ADCk and K parameter maps (referred here as 2D Curvature). Low and high-pass filters were implemented before extracting statistics from signal to evaluate the performance of other implemented feature extraction techniques in relation to simple noise filtering to the data.

In order to utilize all three dimensions of the data, a three-dimensional version of Laws features was implemented (66). Shape feature of surface curvature (11,44) was extracted from 3D mesh after applying marching cubes algorithm to the binary lesion. All 3D feature extractions were preceded by resampling image data into isotropic voxel size. When applicable, all features were calculated from whole prostate, PCa lesions, and whole gland region excluding the lesion. Multiple parameters were applied to feature extraction algorithms where considered suitable. The feature extraction process was applied individually to first scan (**Scan1**), second scan within same day (**Scan2**), and using average feature value of two repetitions (see Fig. 1). In total, ADCk and K parametric maps were processed with 1694 features when statistics and different parameters were considered as separate features. The summary of applied features and number of features in their 18 respective groups is listed in Supporting Information Table S2.

Assessment of the repeatability of radiomics

Short-term repeatability was evaluated for individual radiomic features and machine learning classifiers with ICC, varying between [0..1], where 1 denotes perfectly repeatable individual measurement in relation to variation between subjects and scans, and 0 signifying no repeatability. While ICC values are not directly applicable to other related studies due to differences between subjects in other cohorts, it gives good estimates for comparing between features and between machine learning approaches within the same dataset. To study repeatability of classification between significant and insignificant PCa, the 112 patients were split randomly into 70–30% (as e. g. (44)) development-testing sets stratifying ratio of cases containing at least one lesion with GGG 1, as a trade-off between amount of training data and statistical power in final evaluations with test data. In test dataset 34 cases was considered sufficient to make reasonable performance estimations with unseen data, while rest of the subjects were left to training to make it possible to compare between numerous individual features, while able to apply machine learning. The full data flow of the experiment with radiomics applied to **Scan1** and **Scan2** in multivariate and univariate analysis, with and without help from information from ICC, is shown in Fig. 1.

Four conventionally used feature ranking methods were applied to reveal their effectiveness in terms of selecting features with high classification performance and repeatability:

1. Area under receiver operator characteristic curve (AUC) using leave-pair-out cross-validation (LPO-AUC) procedure to address for potential biases due to sample size (67) for estimating classification potential of individual features for differentiating PCa with GGG 1 vs >1. AUC is conventionally used as a performance metric for raw parameter map values of DWI models (15), giving expected classification potential between low and high groups.

2. Maximum Relevance Minimum Redundancy (MRMR) algorithm using GGG group 1–5 as target. MRMR is conventionally used as a feature pruning method and it considers also the correlation between variables within set of features, minimizing redundancy within chosen feature set.

3. Spearman correlation ρ between feature and GGG groups 1–5, which gives statistical test or non-linear association to all GGG, and used due to non-normality of used features.

4. Wilcoxon rank-sum test for classification of PCa with GGG 1 vs >1, with continuity correction. Like Spearman correlation, Wilcoxon rank-sum test is used as statistical test between low and high GGG groups in non-parametric manner.

Two of the feature ranking methods (LPO-AUC and Wilcoxon rank-sum test) are close to each other due to evaluating binary classification, while the other two (MRMR and Spearman) were applied to evaluate pruning of features when using more detailed information of the target variable GGG. All of the evaluated ranking methods have limitation of addressing signal noise only based on single repetition at hand. We evaluated the effect of refining the ranking method together with ICC values. In addition, we calculated the number of common features shared by selections from **Scan1** and **Scan2** and median(range) of ICC statistics of selected features.

Regularized least-squares (RLS) and ridge regression (68) were used to assess the classification potential of different multivariate radiomic procedures for differentiating PCa with GGG 1 vs >1 using selected features.

Machine learning and feature extraction techniques were implemented with publicly available tools in Python v. 2.7 and v. 3.6, using Meshlab (69) tool in 3D surface processing. Machine learning was implemented using the publicly available RLScore (version 0.8.1, https://github.com/aatapa/RLScore) with AUC (LPO-AUC) (67) when evaluating the number of used features (from 2 to 5), and regularization parameters (from 1.0^{-1} to 1.0^{-6}). Feature ranking criteria and statistical analyses were implemented using R (version 3.5.1, R Foundation for Statistical Computing, Vienna, Austria). Final comparison of radiomics performance was evaluated in testing set (51 lesions) in terms of agreement of their labeling (GGG 1 vs >1) between **Scan1** and **Scan2** (i. e. short-term labeling agreement), and AUC calculated with both scans using trapezoidal rule with 95% confidence interval (DeLong) to describe uncertainty in AUC measurements.

Overview of ROIs and DWI data sets (ADCk and K maps of the training data set) is provided at <<to be added upon acceptance of the manuscript>>. Free public access to post-processing code of most relevant features is provided at <<to be added upon acceptance of the manuscript>>. Imaging data sets are available for non-commercial development pending executed material transfer agreement.

Statistical analysis

We evaluated mutual agreement between feature pruning methods and repeatability of their rankings with Kendall's W (70) having values from 0 (no agreement) to 1 (perfect agreement between rankings). Non-linear monotonic association between repeatability estimates ICC and feature rankings with averaged feature values over repetitions were tested with Spearman's ρ test. We tested performance difference in terms of ICC and ranking values of AUC, MRMR, Spearman and Wilcoxon, with non-parametric Kruskal-Wallis test, followed by Wilcoxon post hoc test to see significance of differences between feature groups where p<3.27e-04 was considered statistically significant.

We also tested for difference of final classification potential AUC when classifier was trained with **Scan1** and **Scan2** data, between using top 10 features suggested by four pruning methods, and by using top 10 features after pre-excluding features not presenting repeatability with ICC>0.8. All tests were implemented with R (version 3.5.1), raw p-values were reported if found statistically significant with level p<0.05, unless otherwise noted.

Results

In total, 170 PCa lesions were present, of those (16%, 28/170), (45%, 77/170), (18%, 31/170), (17%, 29/170) and (3%, 5/170) tumors were to Gleason Grade Group (GGG) 1, 2, 3, 4, and 5, respectively. The training data sets (Fig. 1) consisted of 78 patients with 119 lesions, of those 17, 55, 22, 22 and 3 tumors were to GGG 1, 2, 3, 4, and 5, respectively. The testing data sets consisted of 34 patients with 51 lesions, of those 11, 22, 9, 7 and 2 tumors were to GGG 1, 2, 3, 4, and 5, respectively. The ICC of PCa lesion median intensity of the ADCk and K parameters was 0.786 (95%CI 0.784–0.789) and 0.780 (95%CI 0.778–0.783) in training set, and 0.889 (95%CI 0.887–0.891) and 0.857 (95%CI 0.855–0.860) in the testing set, correspondingly.

Classification Performance and Repeatability of Feature Groups

The classification performance and repeatability, ICC, of all 1694 radiomic features for median feature performance estimate values from **Scan1** and **Scan2** are presented in Fig. 2–3 for ADCk and K, performance of ten highest-ranking features in Fig. 4–5 of each of the 18 feature groups, and individual feature performance in Fig. 6–7. In Fig. 2–7, features with high repeatability but low classification performance in terms of AUC for GGG 1 vs >1 indicate that the feature contains a repeatable signal that is not useful for detection of PCa with GGG 1 vs >1, while these features might potentially still be useful due to their repeatability in other application. In contrast, features with low repeatability, together with high ranking, may contain more uncertainty in the ranking estimate itself. The corresponding plots for AUC, MRMR, Spearman ρ and Wilcoxon rank-sum methods are presented in Supporting Information Figures S1-S4 (ADCk) and S7-S10 (K). Feature group rankings with LPO-AUC and Spearman ρ demonstrated the biggest difference in their ranking of feature groups in comparison to other methods in Fig. 2–5, while differences between rankings were not found to be statistically significant (p>0.05).

There were significant differences between ICC distributions of feature groups (Kruskal-Wallis p≤8.68e-11). A 3D Shape feature and 2D Corner detectors were found to be most repeatable groups in ADCk and K while their ICC difference was not found significant (p>0.05). With all features, 2D Corner detectors outperformed other feature groups in ADCk and K, except 3D Shape (p≤6.19e-05). Similarly, when only top 10 features were considered, 2D Corner detectors outperformed other feature groups (p≤3.0e-4) except five: 3D Shape, Statistics, Statistics (Whole Gland), 2D Wavelet and 2D FFT Band in ADCk, while no difference was found in K.

For performance estimates with all features (Fig. 2–3), in AUC estimates 2D Curvature features outperformed other feature groups except 3D Shape, 2D Corner detectors, 2D Hu and Statistics (ROI Refinement) in ADCk, and in K, 3D Shape and 2D Hu (p≤7.04e-06). Best performing feature group was less prominent with MRMR ranking as in ADCk, 2D Hu had higher scores than 2D Corner detectors and 2D Gabor (p≤2.93e-04), while in K, CoLlaGe features had higher score than 2D Local Binary Pattern, 3D Laws, 3D FFT Band, Statistics, 2D Corner Detector and 2D Gabor (p≤8.14e-06). CoLlaGe was found second best feature group after 3D Shape feature in Wilcoxon test based rankings, with significant difference only to 2D Wavelets and 3D Laws in ADCk, and to 3D Laws in K, while no other major differences were between two best groups to the others. Similar to AUC, 2D Curvature features had highest ranking with Spearman method, with significant difference to all other groups except 3D Shape feature, 2D Corner detectors, 2D Hu, Statistics (ROI Refinement) in ADCk, to all other groups except 3D Shape feature and 2D Hu in K.

Considering only six feature groups with highest ICC (see above), with top 10 features (Fig. 4–5) according to AUC ranking, 2D Corner detectors again outperformed all other feature groups except 3D Shape feature (p≤1.8e-04) with AUC=0.749 (95%CI 0.611..0.887) in ADCk, and all (p≤0.00167) except 3D Shape feature (p=0.154) and 2D Wavelets (p=0.00726) in K. In MRMR estimates, 2D Corner detector was considered significantly better only to Statistics, in ADCk and K (p≤7.15e-05). According to Spearman method rankings, 2D Corner detector was significantly better than Statistics (Whole Gland), 2D Wavelets and 2D FFT Band (p≤1.08e-5) in ADCk, while differences were not significant in K (p≥9.99e-4). Finally, with Wilcoxon performance estimate, 2D Corner detectors were ranked higher than 2D Wavelet, Statistics and Statistics (Whole Gland) (p≤1.6e-04) in ADCk, and Statistics in K (p≤1.7e-04).

Notably, conventionally used statistics from raw intensity values (Statistics, Statistics (Whole Gland), Statistics (ROI Refinements)) and naïve frequency-based filtering approaches (2D FFT Band) were shown to have good repeatability, in comparison to some texture features. Feature group of Statistics (i. e. basic statistics inside lesion), had better

ICC than 2D Local Binary pattern, 2D Haralick, 2D Zernike, 2D Hu, CoLlaGe and 2D Laws features, although not found statistically significant (p≥1.3e-03) in ADCk and significantly better ICC in K than CoLlaGe features (p=4.33e-05). In overall terms, 2D Corner detectors and 3D Laws features provided good repeatability together with four performance metrics.

Repeatability Analysis of Feature Ranking Methods with Individual Features

There was a moderate negative association between ICC and MRMR rankings of **Scan1**, **Scan2**, and averaged features values in ADCk and K with Spearman ρ from −0.579 to −0.494 (all p≤1.08e-104). Correspondingly, rankings of Spearman and AUC had weak association with ρ from −0.358 to −0.158 (all p≤6.310e-11) and ρ from 0.173 to 0.360 (all p≤9.279e-13) in ADCk and K. Wilcoxon feature pruning method had very weak association with ρ from −0.168 to −0.150 (all p≤6.753e-10) in ADCk, while no significant association was found in K (p>0.05).

Three repeatability metrics of agreement between selected features over repetitions, number of common features, and ICC of selected features, are shown for taking best features ranging from top 2 to top 25 features in Fig. 8 for AUC as feature pruning method. All of the three metrics showed improvement when inclusion criteria of high ICC features were applied for AUC, and similarly for MRM, Spearman and Wilcoxon as feature pruning methods (see supporting information figures). MRMR and Wilcoxon feature pruning methods demonstrated significant agreement (Kendall’s W) between **Scan1** and **Scan2** in ADCk (p=0.030 and p= 1.469e-11) and K (p=0.045 and p=1.104e-11) and with Wilcoxon method when only ICC>0.8 was considered (p=0.041).

For averaged feature values, AUC was found to give concordant ranking with Wilcoxon (p=0.036) in ADCk, while other agreements between rankings were not considered statistically significant (p>0.05). Similar to analysis with overall feature groups, among highly repeatable features having also adequate AUC, most represented feature groups in **Scan1** and **Scan2** among top 10 selected features was 3D Laws in ADCk (60%) and in K (51.25%). 2D Corner detector features were most represented feature group when ICC>0.8 inclusion criteria was applied in ADCk (60%) and in K (76.25%). Also with ICC>0.8, 2D Corner detectors was the most represented feature group in all individual feature ranking methods in ADCk and K, except in Wilcoxon method, having 2D FFT Band (50%), 2D Corner detector (20%) and 3D Laws (20%).

Repeatability Analysis of Machine Learning

Repeatability analysis of ADCk parameter alone and combined ADCk and K are given in Table 1 and Table 2. Generally, the four ranking methods gave high rank to features which were fairly repeatable (Table 1, column **C**), while MRMR gave high ranking to features with low repeatability, resulting in selection of feature sets that had poor repeatability values. The final classification performance was not found to be significantly different between the four feature pruning methods (p>0.05). Classification labels (GGG 1 vs >1) varied between classifier trained with **Scan1** and **Scan2 (**column **F),** meaning that classifiers labeled different cases in the test set to positive and negative depending on which repetition was used in training. There was a small variation between AUC estimates in the test **Scan1** and

Scan2 with largest difference of 0.07 with LPO-AUC and Spearman ρ in Table 2, while the differences were not found to be statistically significant (p>0.05). When additional information from repeatability was included (pre-inclusion of only features with high repeatability ICC>0.8), the selected feature sets were generally more consistent and labeling of classifiers trained with either **Scan1** and **Scan2** were in better agreement, particularly in Spearman ρ method.

The performance of machine learning when feature extraction signal was pooled together from both parameters ADCk and K (top 5 from ADCk and top 5 from K) of kurtosis function is shown in Table 2. The difference in AUC between **Scan1** and **Scan2** of testing set, suggesting that the four applied feature ranking methods still contained unaddressed noise which further propagated as fitting model with a suboptimal group of features. In comparison to reference method of taking ADCk or K median value (0.63 and 0.68) inside the lesion, individual features as representatives of univariate models improved the classification, and RLS classifier gave performance estimate up to 0.77, and same feature sets resulted up to 0.75 for classification of GGG 1 and 2 vs >2.

Discussion

Radiomics including textures and shape feature, and ML methods hold promise for improved patient's care by introducing objective optimized medical image evaluation. Performance of radiomics is often evaluated in terms of AUC. Despite an exponentially increasing number of studies using radiomics and ML in medical imaging, only a handful of these studies evaluated the repeatability and reproducible of these methods (8), and uncertainty of given AUC classification estimate is rarely given. Although in the current study the found differences in machine learning classifications were not found statistically significant, we observed improvement in both reliability and performance of final classifications between GGG 1 and >1, when repeatability of features was taken into account in DWI radiomic feature selection (Table 1) and in classifier training (Table 2), having AUC=0.77 (0.64..0.89) similar to study by Starmans et al. (71) with 40 subjects. Further, we observed significant differences between repeatability values of the feature groups, suggesting of using most repeatable features.

In the current study, we evaluated short-term repeatability and diagnostic performance of commonly used radiomics together with machine learning for PCa classification using DWI obtained with 12 b-values up to 2000 s/mm^2. Some of the conventional radiomics such as texture methods demonstrated high AUC but low repeatability, stressing the fact that high classification potential in training set does not necessarily mean good overall performance, specifically for classifying PCa between GGG 1 vs >1. Therefore, features with low repeatability but high classification performance for PCa classification should be considered with caution, as the feature may turn out to have poor short-term repeatability, while high short-term repeatability is needed for practical application of them in non-invasive PCa detection, characterization, therapy planning, and therapy monitory (5).

A number of factors affect the performance of radiomic features. For example, statistical descriptors can be particularly sensitive to the intensity of the image. Texture features of

DWI, among other radiomic features, have already demonstrated some promise for PCa detection and management, e. g. (23,24). However, a large variation in radiomics performance has been observed in multi-institutional studies (5). Texture features find texture patterns from the image and are less dependent on the intensity than statistical descriptors, but are sensitive to voxel level variations. Shape features (11,27) can rely partially or solely on the human delineation, and as such are expected to be more robust against voxel level intensity variations within and between images. In agreement of our findings, surface-to-area ratio has been shown to perform well for the detection of clinically significant PCa (11). In a small study of 8 women with cervical cancer who underwent test–retest MRI, shape features and topological features demonstrated high repeatability (72) similar to the current study. Surface-to-area ratio has limitation of not incorporating information from absolute values of ADCk and K parameter maps which are known to correlate with PCa (46). However, their potential use may partially be in combination with other radiomic features.

Prostate MRI radiomic features as imaging-based markers are typically evaluated based on AUC values for PCa detection/characterization. Repeatability (same scanner/protocol) and reproducibility (different scanner/protocol) of radiomics and ML for prostate DWI in the same patients have not been evaluated so far. A study by Chirra et al. (73) evaluated external reproducibility of radiomics derived from prostate T2-weighted imaging but repeatability and reproducibility of radiomics and ML for prostate DWI in the same patients with PCa has not been evaluated. We found that the repeatability of the selected features was dependent on the applied feature ranking method. This may be due to all of the evaluated selection techniques focusing on metrics with which repeatability varies independently. We speculate that the within-subject variability propagates to the feature selection metrics, causing subsequently changes in the feature ranking order, depending on robustness of ranking methods. As the evaluated methods use only one repetition, they can potentially provide only indirect assessments of within-subject variations. We did not find any of the evaluated performance estimation methods to be considered as candidates for estimating repeatability, suggesting that to address repeatability, other means would need to be used, such as repeated scans. While we stress importance of direct measurements for repeatability, we suggest using Wilcoxon rank-sum test as feature selection criteria due to most consistent feature selections between repetitions. With our evaluated features using more top-ranking features caused the overall repeatability of feature selection to diminish. Thus, no warrant is given for blindly taking, for example, the top 10 of the best ranked features without knowledge of the repeatability of the feature selection process itself (Fig. 5). Careful optimization of radiomic parameters and feature ranking methods are needed to achieve both high diagnostic performance for PCa characterization as well as reliable labeling for clinical use.

This study has multiple limitations. We did not correct for possible correlation between multiple tumors in individual PCa patients. In 46 (41%, 46/112) patients, more than one PCa lesion was present. We focused on short-term repeatability of features extracted from the data and derived machine learning for practical application of separating GGG groups from each other. While most of the features have intuitive biophysical explanation such as shape or larger scale density variation in the tissue (textures, gradient methods) we did not perform separate evaluations for how well the radiomic features would correspond to histology. It is to be noted that repeatability and reproducibility of radiomic feature are additional requirements to good classification performance together with correspondence to histology. We addressed classification between significant and insignificant PCa, due to significant implication on patient's management. Classification to other GGG is left for future work. Modeling of PCa DWI signal decay is most commonly performed using the mono-exponential model. However, kurtosis function has higher information content (fitting quality), similar repeatability, and robustness against noise (45–47), thus kurtosis function was used in the current study (39). Future studies are needed to evaluate the performance of radiomics derived using different function/models for prostate DWI. While we used kurtosis function to addresses the deviation of DWI signal decay at high b-values, it is important to stress that our results may still be influenced by the DWI acquisition parameters, such as TE, b-values (12 b-values in the range from 0 to 2000 s/mm^2) and diffusion time (20.3 ms).

While we consider the applied features in this study to be a good representation of techniques given in related literature with all main types of features, the list does not consider all possibilities, and only one classifier was used in evaluations. Prostate cancer lesions on DWI data were manually delineated by one research fellow working in consensus with the genitourinary pathologist, using whole mount prostatectomy sections as "ground truth". Other modalities besides DWI were not considered and it is left for future studies to explore effects of additional MRI sequences. Future studies are needed to evaluate the performance of calculated features using fully automatic tools for lesion delineation, as semi or fully automatic tool for ROI definition could be used in conjunction with feature selection methods in radiomics. Noting that only well repeatable features are to be expected to have link to underlying pathology, more detailed biophysical explanation of those features is left for future study. Although to best of our knowledge, the number of repeated DWI scans and radiomic features was the largest for evaluating short-term repeatability of prostate MRI derived features (38), it could be argued that the study is still limited by its sample size.

Conclusion

In this study, we have shown that: 1. Only a fraction of radiomics from ADCk and K had high repeatability, ICC(3,1)>0.8; 2. Shape and 2D Corner detector-based features were among the most represented features with high repeatability, ICC(3,1)>0.75 and high classification performance, AUC>0.70; 3. The applied feature selection method did have an a major effect on repeatability, ICC(3,1), and final classification performance, AUC, while the difference did not reach the level of statistical significance; 4. Although none of the feature selection methods selected the most repeatable features, Wilcoxon rank-sum test as feature selection criteria was found to be most consistent between repetitions. We demonstrated that radiomic features and ML methods should be addressed with caution if only single scan data are available and repeatability is unknown, and that repeatability improved performance in unseen data in terms of both absolute performance and stability of classifications between test and retest scan of unseen data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was financially supported by grants from Sigrid Jusélius Foundation (HM, HA, IJ), Finnish Cultural Foundation and Orion Pharma Research Foundation (HM), clinical Researcher Funding from the Academy of Finland (PT), TYKS-SAPA Research funds (HM, PT, OE, MP, JS, JPB, HA, IJ), Finnish Cancer Society (IJ), Instrumentarium Reseseach Foundation (IJ, HA), DOD United States W81XWH-18–1-0524 (RS, AM) and grants from NIH (1U24CA199374–01, R01CA202752–01A1, R01CA208236–01A1, R01 CA216579–01A1, R01 CA220581–01A1, 1U01 CA239055–01), NCRR (C06 RR12463–01), VA Merit Review Award IBX004121A and DOD United States (W81XWH-15–1-0558, W81XWH-16–1-0329, W81XWH-18–1-0524) (AM). Funding from Instrumentarium Science Foundation, Sigrid Jusélius Foundation and Turku University Hospital, TYKS-SAPA research funds were used to cover the cost of MRI examinations.

We thank Anitta Entonen (Turku University Hospital, Turku Finland) for her help with patient enrollment and assistance during imaging of PCa patients, and Jaakko Liippo (Turku University Hospital, Turku, Finland) for his help in scanning the histopathological slides, and Ileana Montoya Perez (Turku University, Turku, Finland) and Timo Laitinen for their help with establishing and running the study server.

References

1. Wang Y-XJ, Ng CK. The impact of quantitative imaging in medicine and surgery: Charting our course for the future. Quant Imaging Med Surg 2011 12;1(1):1–3. [PubMed: 23256048]
2. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6. [PubMed: 22257792]
3. Pinker K, Shitano F, Sala E, Do RK, Young RJ, Wibmer AG, et al. Background, current role, and potential applications of radiogenomics. J Magn Reson Imaging. 2018 3;47(3):604–20. [PubMed: 29095543]
4. Forghani R, Savadjiev P, Chatterjee A, Muthukrishnan N, Reinhold C, Forghani B. Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology. Comput Struct Biotechnol J [Internet]. 2019;17:995–1008. Available from: 10.1016/j.csbj.2019.07.001
5. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019;69(2):127–57. [PubMed: 30720861]
6. Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks. 2006;19(4):408–15. [PubMed: 16483741]
7. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol [Internet]. 2019;110:12–22. Available from: 10.1016/j.jclinepi.2019.02.004
8. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018 11;102(4):1143–58. [PubMed: 30170872]
9. Zhu X, Wu X. Class Noise vs. Attribute Noise: A Quantitative Study. Artif Intell Rev 2004;22(3):177–210.
10. Sun Y, Reynolds HM, Parameswaran B, Wraith D, Finnegan ME, Williams S, et al. Multiparametric MRI and radiomics in prostate cancer: a review. Australas Phys Eng Sci Med 2019 3;42(1):3–25. [PubMed: 30762223]
11. Cuocolo R, Stanzione A, Ponsiglione A, Romeo V, Verde F, Creta M, et al. Clinically Significant Prostate Cancer Detection on MRI: a Radiomic Shape Features Study. Eur J Radiol [Internet]. 2019;116(March):144–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0720048X19301664
12. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solórzano G, Erho N, et al. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res 2016;5(4):432–47. [PubMed: 29188191]
13. Sidhu HS, Benigno S, Ganeshan B, Dikaios N, Johnston EW, Allen C, et al. Textural analysis of multiparametric MRI detects transition zone prostate cancer. Eur Radiol 2017;27(6):2348–58. [PubMed: 27620864]

14. Kim S, Decarlo L, Cho GY, Jensen JH, Sodickson DK, Moy L, et al. Interstitial fluid pressure correlates with intravoxel incoherent motion imaging metrics in a mouse mammary carcinoma model. Vol. 25, NMR Biomed p. 787–94.

15. Khalvati F, Wong A, Haider MA. Automated Prostate Cancer Detection on Multi-parametric MR imaging via Texture Analysis. BMC Med Imaging. 2015;15(1):27. [PubMed: 26242589]

16. Tiwari P, Kurhanewicz J, Madabhushi A. Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. Med Image Anal [Internet]. 2013;17(2):219–35. Available from: 10.1016/j.media.2012.10.004

17. Merisaari H, Shiradkar R, Toivonen J, Hiremath A, Khorrami M, Perez Montoya I, et al. Repeatability of radiomic features for prostate cancer diffusion weighted imaging obtained using b-values up to 2000 s/mm2. In: ISMRM 27th Annual Meeting & Exhibition, Montreal, QC, Canada 2019 p. 4472.

18. Madabhushi A, Feldman MD, Metaxas DN, Tomaszeweski J, Chute D. Automated detection of prostatic adenocarcinoma from high-resolution Ex vivo MRI. IEEE Trans Med Imaging. 2005;24(12):1611–25. [PubMed: 16350920]

19. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 11;68(6):394–424. [PubMed: 30207593]

20. Jambor I. Optimization of prostate MRI acquisition and post-processing protocol: a pictorial review with access to acquisition protocols. Acta Radiol Open [Internet]. 2017;6(12):205846011774557. Available from: http://journals.sagepub.com/doi/10.1177/2058460117745574

21. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med [Internet]. 2018 5;378(19):1767–77. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1801993

22. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40(2):244–52. [PubMed: 26492179]

23. Shiradkar R, Ghose S, Jambor I, Taimen P, Ettala O, Purysko AS, et al. Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: Preliminary findings. J Magn Reson Imaging [Internet]. 2018 5; Available from: http://www.ncbi.nlm.nih.gov/pubmed/29734484

24. Algohary A, Viswanath S, Shiradkar R, Ghose S, Pahwa S, Moses D, et al. Radiomic features on MRI enable risk categorization of prostate cancer patients on active surveillance: Preliminary findings. J Magn Reson Imaging [Internet]. 2018 9;48(3):818–28. Available from: http://doi.wiley.com/10.1002/jmri.25983

25. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med [Internet]. 2019;25(1):30–6. Available from: 10.1038/s41591-018-0307-0

26. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: A review of statistical analysis of repeat data sets. Transl Oncol 2009;2(4):231–5. [PubMed: 19956383]

27. Schwier M, Van Griethu J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci Rep 2019;9(1):9441. [PubMed: 31263116]

28. Gibbs P, Pickles MD, Sreenivas M, Knowles a, Turnbull LW. Repeatability of Diffusion Imaging of the Prostate at 3T. 2005;13(2004):2005.

29. Merisaari H, Jambor I. Optimization of b-value distribution for four mathematical models of prostate cancer diffusion-weighted imaging using b values up to 2000 s/mm2: Simulation and repeatability study. Magn Reson Med 2015;73(5):1954–69. [PubMed: 25045885]

30. Jambor I, Merisaari H, Taimen P, Boström P, Minn H, Pesola M, et al. Evaluation of different mathematical models for diffusion-weighted imaging of normal prostate and prostate cancer using high b-values: A repeatability study. Magn Reson Med 2015;73(5):1988–98. [PubMed: 25046482]

31. Toivonen J, Merisaari H, Pesola M, Taimen P, Bostr PJ, Pahikkala T, et al. Mathematical Models for Diffusion-Weighted Imaging of Prostate Cancer Using b Values up to 2000 s / mm 2 : Correlation with Gleason Score and Repeatability of Region of Interest Analysis. 2015;1124(September 2014):1116–24.

32. Johnston EW, Bonet-Carne E, Ferizi U, Yvernault B, Pye H, Patel D, et al. VERDICT MRI for Prostate Cancer : Intracellular Volume Fraction versus Apparent Diffusion Coefficient. Radiology. 2019;291(2):391–7. [PubMed: 30938627]

33. Sadinski M, Medved M, Karademir I, Wang S, Peng Y, Jiang Y, et al. Short-term reproducibility of apparent diffusion coefficient estimated from diffusion-weighted MRI of the prostate. Abdom Imaging. 2015;40(7):2523–8. [PubMed: 25805558]

34. Malyarenko DI, Newitt D, Wilmes LJ, Tudorica A, Helmer KG, Arlinghaus LR, et al. Demonstration of nonlinearity bias in the measurement of the apparent diffusion coefficient in multicenter trials. Magn Reson Med 2016;75(3):1312–23. [PubMed: 25940607]

35. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;

36. Fedorov A, Vangel MG, Tempany CM, Fennessy FM. Multiparametric Magnetic Resonance Imaging of the Prostate Repeatability of Volume and Apparent Diffusion Coefficient Quantification. Invest Radiol 2017;52(9):538–46. [PubMed: 28463931]

37. Mehta S, Bastero-Caballero RF, Sun Y, Zhu R, Murphy DK, Hardas B, et al. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. Stat Med 2018;37(18):2734–52. [PubMed: 29707825]

38. Fedorov A, Schwier M, Clunie D, Herz C, Pieper S, Kikinis R, et al. Data descriptor: An annotated test-retest collection of prostate multiparametric mri. Sci Data [Internet]. 2018;5:180281. Available from: 10.1038/sdata.2018.281

39. Jensen JH, Helpern J a, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med [Internet]. 2005 6 [cited 2014 Jan 23];53(6):1432–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15906300

40. Toivonen J, Merisaari H, Pesola M, Taimen P, Bostrom PJ, Pahikkala T, et al. Mathematical models for diffusion-weighted imaging of prostate cancer using b values up to 2000 s/mm(2) : correlation with Gleason score and repeatability of region of interest analysis. Vol. 74, Magn Reson.Med p. 1116–24.

41. Jambor I, Merisaari H, Taimen P, Bostrom P, Minn H, Pesola M, et al. Evaluation of different mathematical models for diffusion-weighted imaging of normal prostate and prostate cancer using high b-values: a repeatability study. Vol. 73, Magn Reson.Med p. 1988–98.

42. Peng H, Long F, Ding C. Feature selection based on mutual information. IEEE Trans Pattern Anal Mach Intell 2015;27(8):1226–38.

43. Kwak JT, Xu S, Wood BJ, Turkbey B, Choyke PL, Pinto PA, et al. Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging. Vol. 42, Med.Phys p. 2368–78.

44. Chen T, Li M, Gu Y, Zhang Y, Yang S, Wei C, et al. Prostate Cancer Differentiation and Aggressiveness: Assessment With a Radiomic-Based Model vs. PI-RADS v2. J Magn Reson Imaging. 2019;49(3):875–84. [PubMed: 30230108]

45. Merisaari H, Toivonen J, Pesola M, Taimen P, Boström PJ, Pahikkala T, et al. Diffusion-weighted imaging of prostate cancer: Effect of b-value distribution on repeatability and cancer characterization. Magn Reson Imaging. 2015;33(10):1212–8. [PubMed: 26220861]

46. Toivonen J, Merisaari H, Pesola M, Taimen P, Boström PJ, Pahikkala T, et al. Mathematical models for diffusion-weighted imaging of prostate cancer using b values up to 2000 s/mm2: Correlation with Gleason score and repeatability of region of interest analysis. Magn Reson Med 2015;74(4):1116–24. [PubMed: 25329932]

47. Jambor I, Merisaari H, Taimen P, Boström P, Minn H, Pesola M, et al. Evaluation of different mathematical models for diffusion-weighted imaging of normal prostate and prostate cancer using high b-values: A repeatability study. Magn Reson Med [Internet]. 2015 5;73(5):1988–98. Available from: http://doi.wiley.com/10.1002/mrm.25323

48. Jambor I, Pesola M, Merisaari H, Taimen P, Bostrom PJ, Liimatainen T, et al. Relaxation along fictitious field, diffusion-weighted imaging, and T2 mapping of prostate cancer: Prediction of cancer aggressiveness. Vol. 75, Magn Reson.Med p. 2130–40.

49. Jambor I, Pesola M, Taimen P, Merisaari H, Boström PJ, Minn H, et al. Rotating frame relaxation imaging of prostate cancer: Repeatability, cancer detection, and Gleason score prediction. Magn Reson Med 2016;75(1):337–44. [PubMed: 25733132]

50. Kahkonen E, Jambor I, Kemppainen J, Lehtio K, Gronroos TJ, Kuisma A, et al. In Vivo Imaging of Prostate Cancer Using [68Ga]-Labeled Bombesin Analog BAY86–7548. Clin.Cancer Res p.

51. Tamura C, Shinmoto H, Soga S, Okamura T, Sato H, Okuaki T, et al. Diffusion kurtosis imaging study of prostate cancer: Preliminary findings. J Magn Reson Imaging. 2014;40(3):723–9. [PubMed: 24924835]

52. Quentin M, Pentang G, Schimmöller L, Kott O, Müller-Lutz A, Blondin D, et al. Feasibility of diffusional kurtosis tensor imaging in prostate MRI for the assessment of prostate cancer: Preliminary results. Magn Reson Imaging [Internet]. 2014;32(7):880–5. Available from: 10.1016/j.mri.2014.04.005

53. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Vol. 53, Magn Reson.Med 2005 p. 1432–40.

54. Shanno DF. On broyden-fletcher-goldfarb-shanno method. J Optim Theory Appl 1985;46(1):87–94.

55. Armi L, Fekri-Ershad S. Texture image analysis and texture classification methods - A review. Int Online J Image Process Pattern Recognit [Internet]. 2019;2(1):1–29. Available from: http://arxiv.org/abs/1904.06554

56. Sobel I, Feldman G. A 3×3 isotropic gradient operator for image processing, presented at a talk at the Stanford Artificial Project. attern Classif Scene Anal 1968;271–272.

57. Kirsch RA. Computer determination of the constituent structure of biological image. Comput Biomed Res 1971;328:315–28.

58. Teague MR. Image analysis via the general theory of moments*. J Opt Soc Am 1980 8;70(8):920.

59. Institution of Electrical Engineers. D. The journal of the Institution of Electrical Engineers. Part III, Radio and communication engineering, including the Proceedings of the Radio Section of the Institution. Vol. 93, Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering. IET Digital Library; 1946 429–441 p.

60. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern [Internet]. 1973 11;SMC-3(6):610–21. Available from: http://ieeexplore.ieee.org/document/4309314/

61. Prasanna P, Tiwari P, Madabhushi A. Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe): A new radiomics descriptor. Sci Rep [Internet]. 2016;6:37241. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27872484

62. Ginsburg SB, Viswanath SE, Bloch BN, Rofsky NM, Genega EM, Lenkinski RE, et al. Novel PCA-VIP scheme for ranking MRI protocols and identifying computer-extracted MRI measurements associated with central gland and peripheral zone prostate tumors. J Magn Reson Imaging. 2015;41(5):1383–93. [PubMed: 24943647]

63. Laws KI. Texture energy measures. In: DARPA Image Understanding Worlkshop 1979 p. 47–51.

64. Huan T, Sivachenko AY, Harrison SH, Chen JY. ProteoLens: a visual analytic tool for multi-scale database-driven biological network data mining. BMC Bioinformatics [Internet]. 2008 1 [cited 2014 Jan 23];9 Suppl 9:S5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2537576&tool=pmcentrez&rendertype=abstract

65. Harris CG and Stephens M. A combined corner and edge detector. In: Alvey vision conference 1988 p. 10–5244.

66. Suzuki MT, Yaginuma Y. A solid texture analysis based on three-dimensional convolution kernels. Videometrics IX. 2007;6491(January 2007):64910W.

67. Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. A comparison of AUC estimators in small-sample studies. In: Machine learning in systems biology. 2009 p. 3–13.

68. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970 2;12(1):55–67.

69. Cignoni P, Callieri M, Corsini M, Dellepiane M, Ganovelli F, Ranzuglia G. Meshlab: an open-source mesh processing tool. In: Eurographics Italian chapter conference 2008 p. 129–36.

70. Willinek W a, Gieseke J, Kukuk GM, Nelles M, König R, Morakkabati-Spitz N, et al. Dual-source parallel radiofrequency excitation body MR imaging compared with standard MR imaging at 3.0 T: initial clinical experience. Radiology. 2010;256(3):966–75. [PubMed: 20720078]

71. Starmans MP, Niessen WJ, Schoots I, Klein S, Veenland JF. Classification Of Prostate Cancer: High Grade Versus Low Grade Using A Radiomics Approach. In: IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) 2019 p. 1319–22.

72. Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol 2019;135:107–14. [PubMed: 31015155]

73. Chirra P, Leo P, Yim M, Bloch BN, Rastinehad AR, Purysko A, et al. Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI. J Med Imaging. 2019;6(02):1.

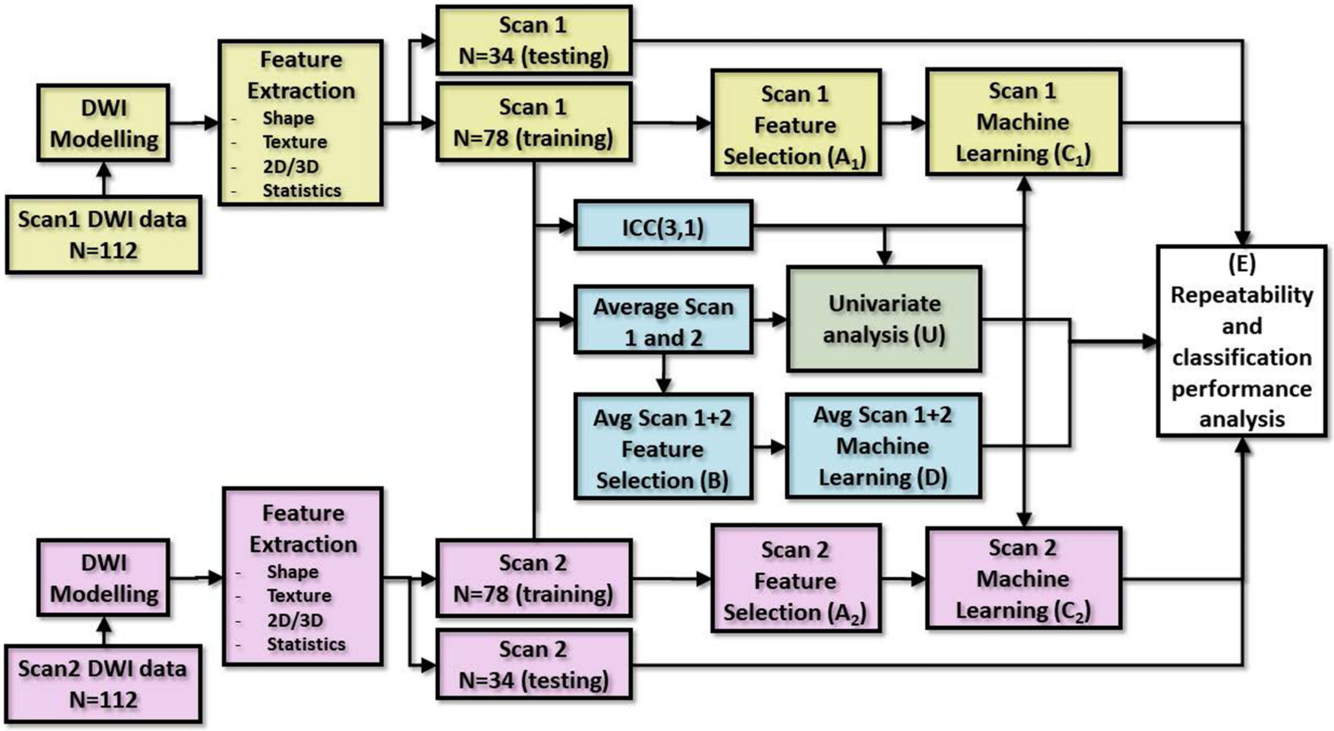

Fig. 1:

Data flow of repeatability analysis for machine learning application using diffusion weighted imaging data of 112 patients with prostate cancer. The process for **Scan1** and **Scan2** is identical apart from data from acquisition time, containing DWI modelling, feature extraction and data split to training and testing sets. (**A$_1$**) and (**A$_2$**): Selection of features based on single repetition approach with **Scan1** and **Scan2**, respectively. (**B**): Selection of features using average feature values of **Scan1** and **Scan2**. (**C$_1$**), (**C$_2$**) and **(D)**: Machine learning of classifier with single repetitions data **Scan1**, **Scan2**, and average of repetitions, respectively. (**E**): Final analysis of stability of trained classifiers and stability of final classifications. (**U**): Results of machine learning are compared to univariate analysis, including conventional DWI metrics.

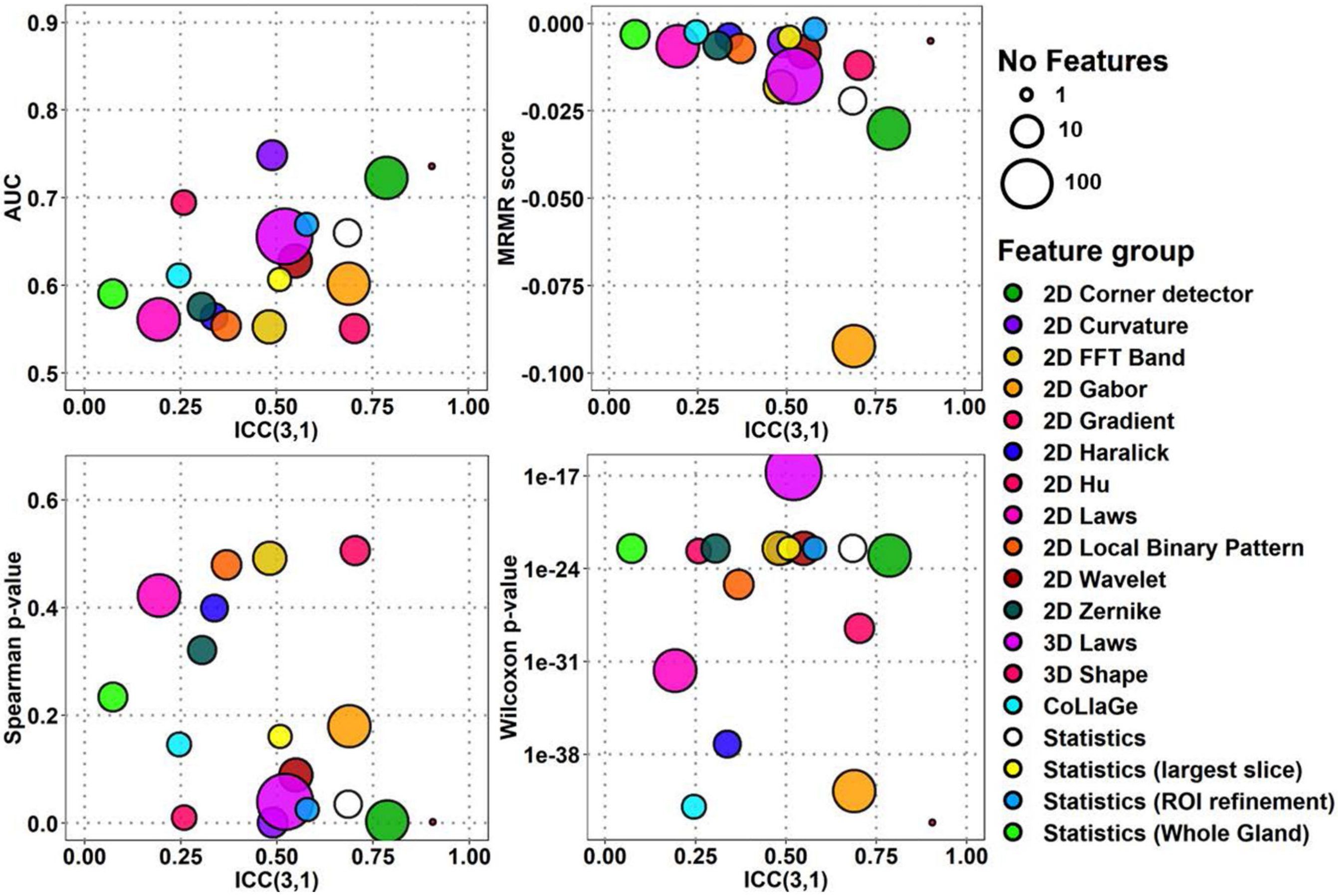

Fig. 2:

Overall feature extraction scores of 1694 features in 18 feature groups, with four ranking methods and repeatability for ADCk (Apparent Diffusion Coefficient) of kurtosis function.

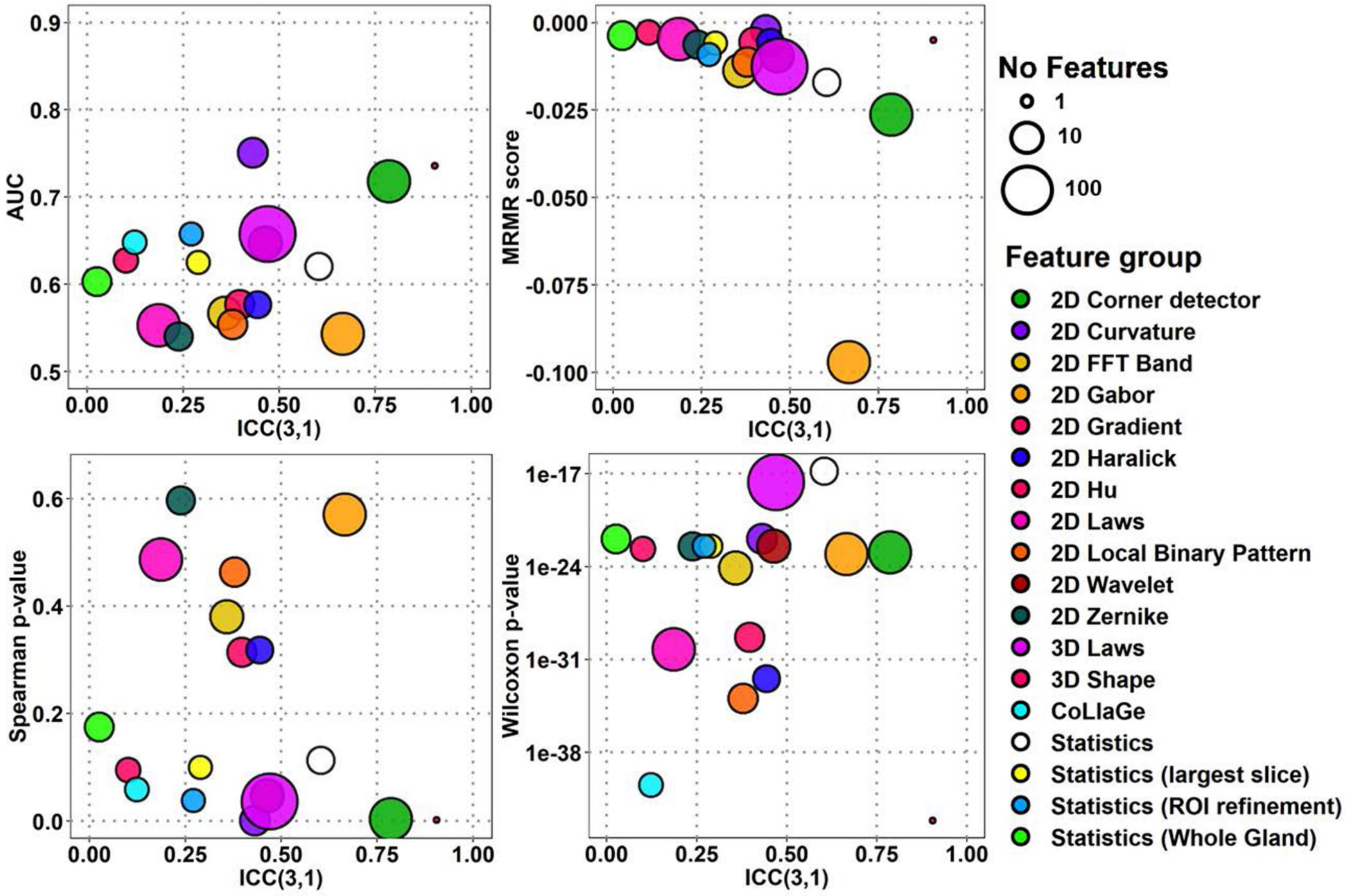

Fig. 3:

Overall feature extraction scores of 1694 features in 18 feature groups, with four ranking methods and repeatability for K (kurtosis) of kurtosis function.

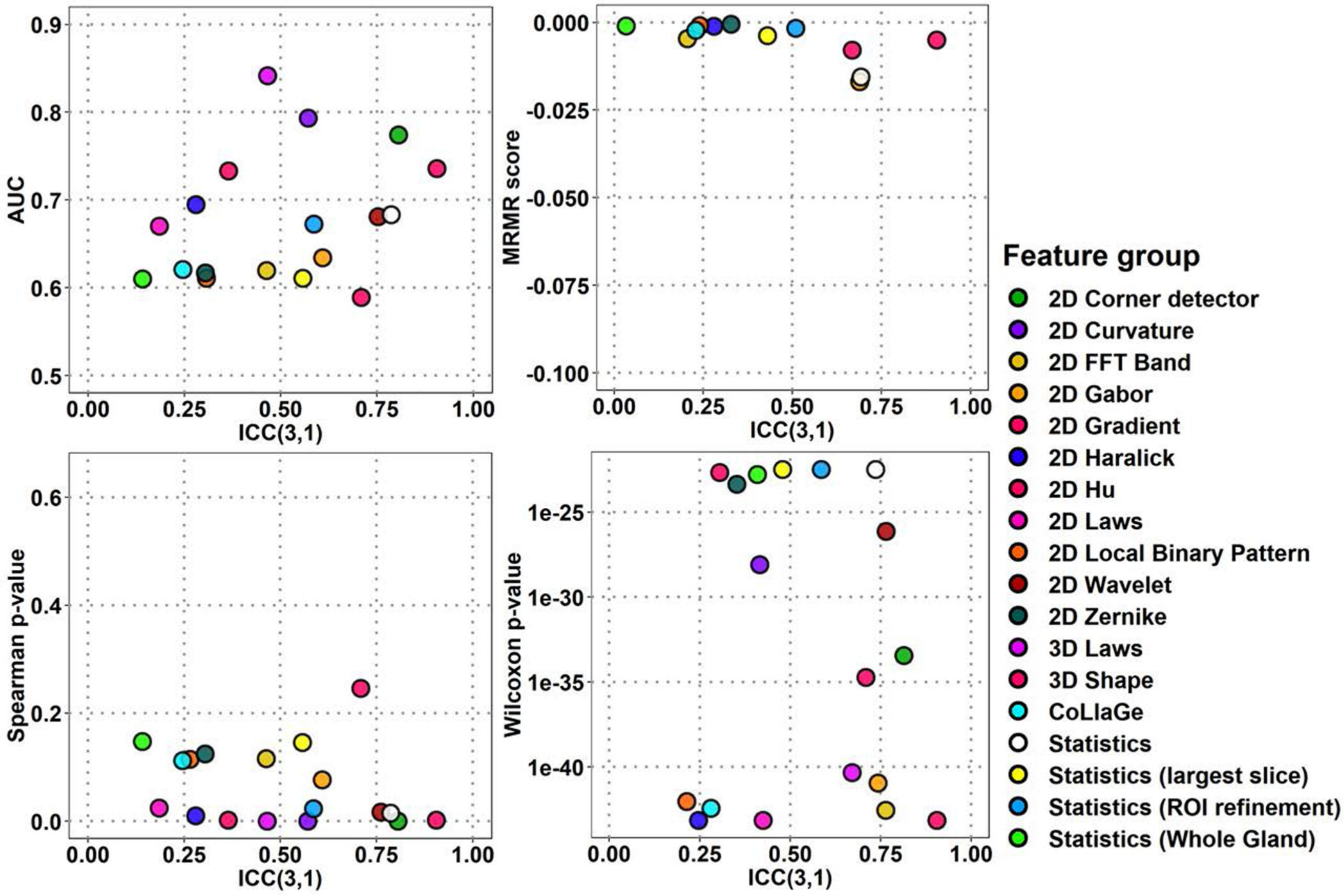

Fig. 4:

Feature extraction rankings with four ranking methods and repeatability for top 10 features in each of 18 feature groups for ADCk (Apparent Diffusion Coefficient) of kurtosis function.

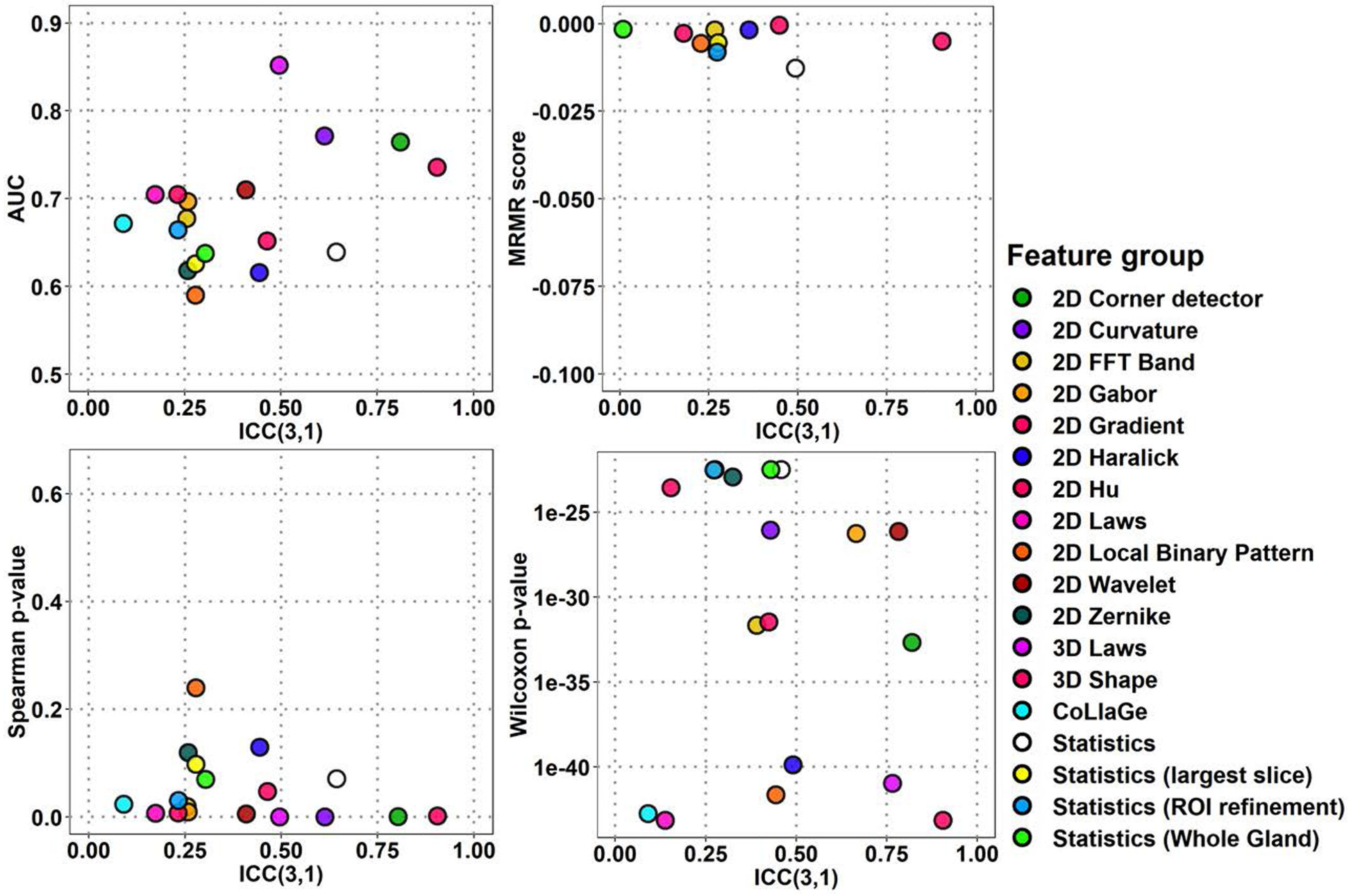

Fig. 5:

Feature extraction rankings with four ranking methods and repeatability for top 10 features in each of 18 feature groups for K (kurtosity) of kurtosis function.

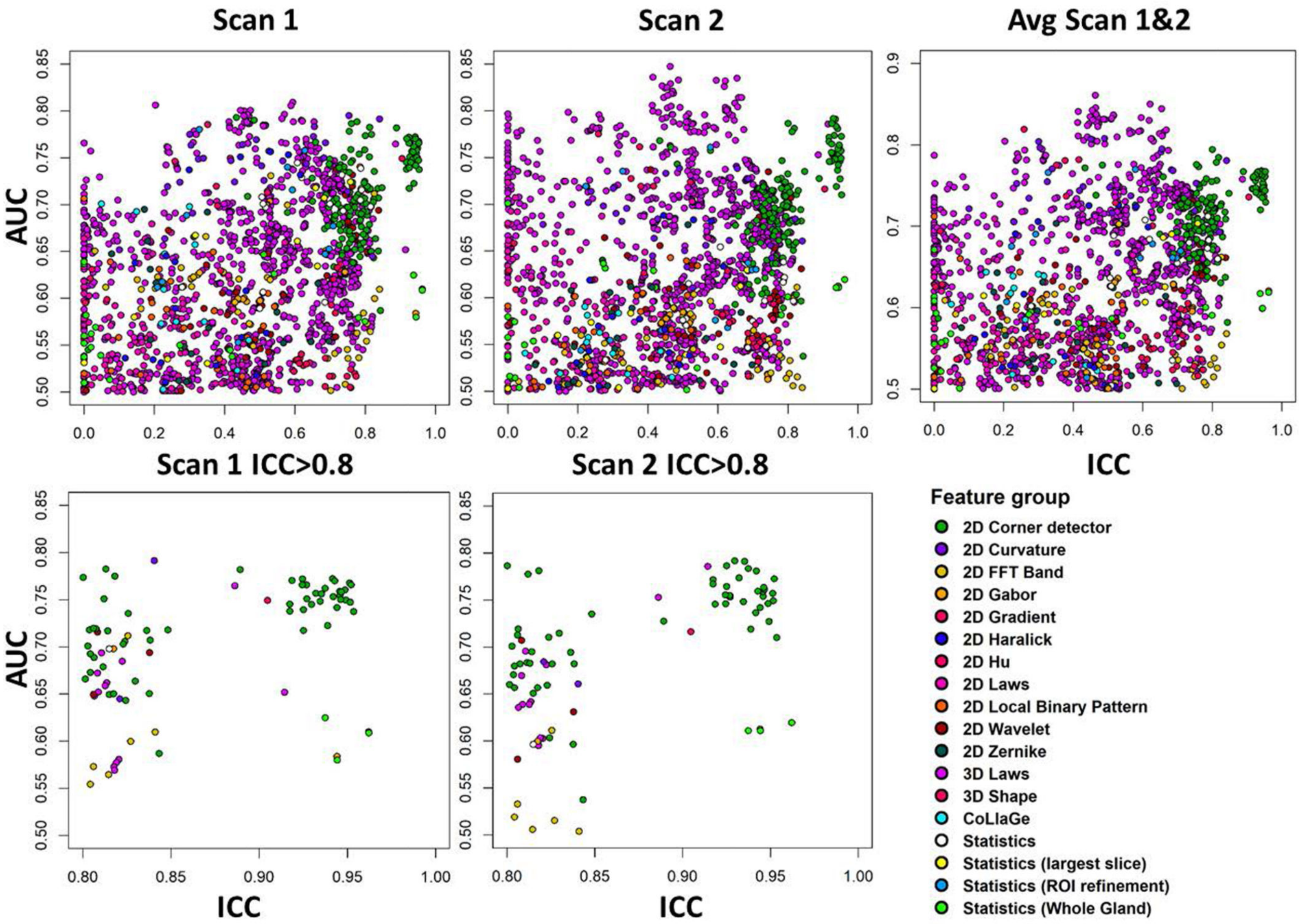

Fig. 6:

Classification performance for prostate cancer classification repeatability for 1694 features derived from ADCk (Apparent Diffusion Coefficient) of kurtosis function, using the area under receiver operator characteristic curve (AUC). The scatterplots show repeatability of radiomic features (x-axis) with respect to AUC (y-axis), for all features (top), and features having high repeatability only (ICC(3,1)>0.8, bottom). Measurements are from single DWI scan (**Scan1** and **Scan2**) and when radiomic feature values are averaged (**Avg Scan 1&2**).

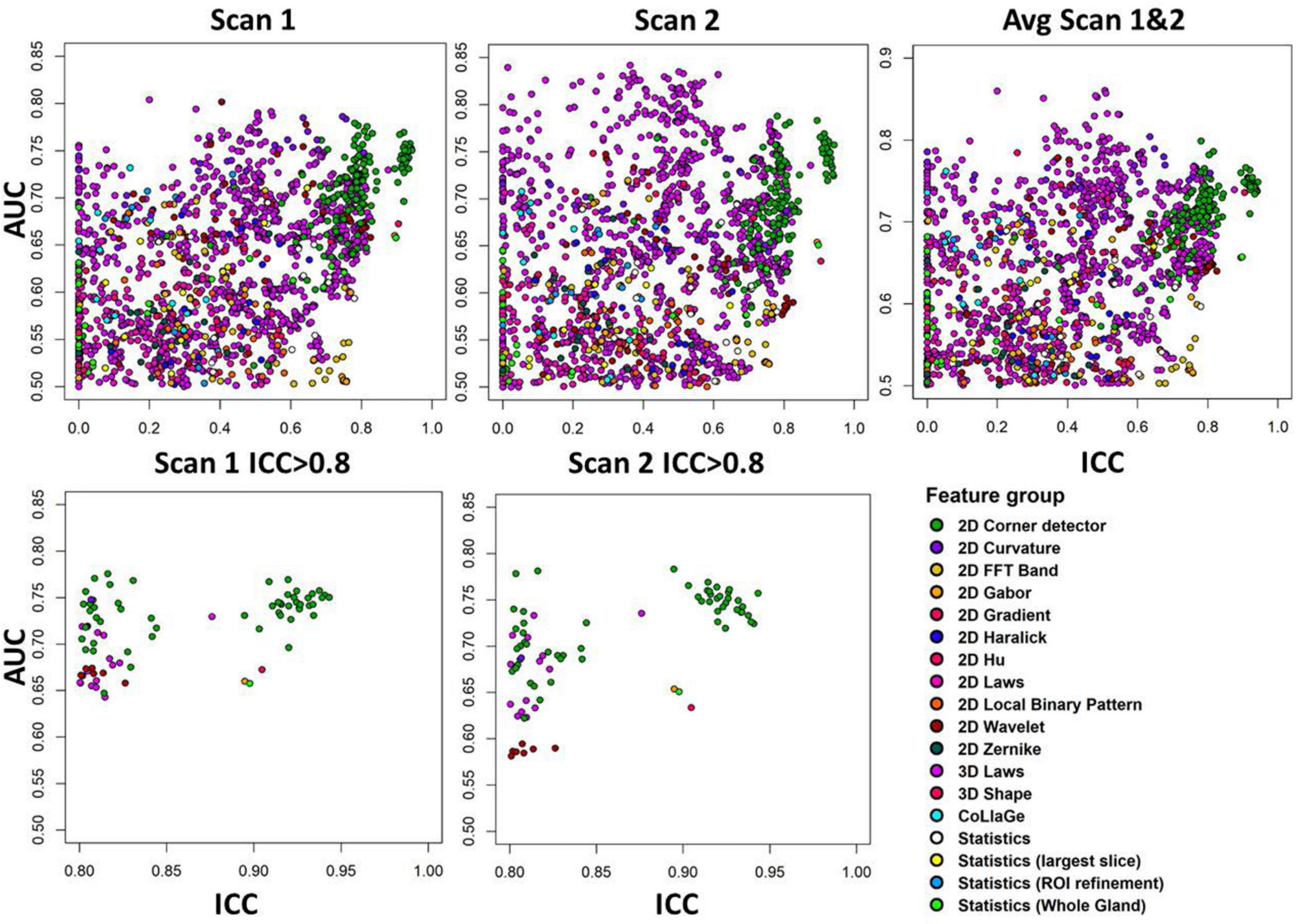

Fig. 7:

Classification performance for prostate cancer classification repeatability for 1694 features derived from K (kurtosity) of kurtosis function, using the area under receiver operator characteristic curve (AUC). The scatterplots show repeatability of radiomic features (x-axis) with respect to AUC (y-axis), for all features (top), and features having high repeatability only (ICC(3,1)>0.8, bottom). Measurements are from single DWI scan (**Scan1** and **Scan2**) and when radiomic feature values are averaged (**Avg Scan 1&2**).

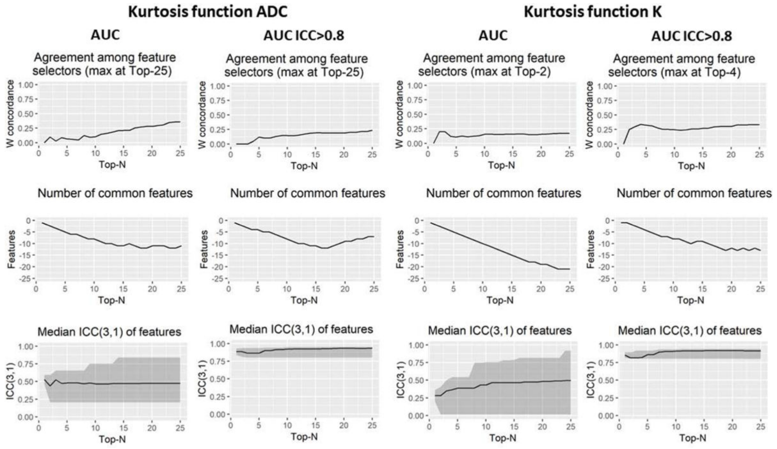

Fig. 8:

AUC feature selection associations with repeatability using ADC (Apparent Diffusion Coefficient) and K (kurtosity) of kurtosis function with and without pre-selection of highly repeatable (ICC(3,1)>0.8) features.

Table 1:

Stability of machine learning results calculated independently in test (**Scan1**) and retest (**Scan2**) scans. (**A**) Feature pruning method. (**B**) Selection of 10 best ranking features according to pruning method. (**C**) Median and range of repeatability of features used in classifier training. **(D)** Agreement of binary classifications from classifiers trained with **Scan1** and **Scan2** data**. (E)** Area Under Receiver operator characteristic curve (AUC) when classifications are evaluated with **Scan1** of test set. **(F)** AUC with **Scan2**. (G) Pooled estimate over **Scan1** and **Scan2** for AUC.

Pruning method (A)	Pruned features (B)	ICC median (range) (C)	% of same labels (D)	AUC (95% CI) Scan1 (E)	AUC (95% CI) Scan 2 (F)	AUC (95% CI) Scan 1 and 2 (G)

Wilcoxon rank-sum test	ADCk Top 10	0.438(0.000..0.953)
		0.280(0.000..0.693)
			80.4	0.682 (0.490..0.873)	0.495 (0.261..0.730)	0.586 (0.436..0.737)

MRMR	ADCk Top 10	0.296(0.000..0.813)
		0.233(0.000..0.621)
			70.6	0.673 (0.469..0.876)	0.630 (0.438..0.821)	0.649 (0.514..0.785)

Spearman ρ	ADCk Top 10	0.471(0.445..0.754)
		0.476(0.414..0.654)
			70.6	0.743 (0.558..0.928)	0.645 (0.465..0.826)	0.691 (0.563..0.819)

AUC	ADCk Top 10	0.471(0.445..0.754)
		0.476(0.414..0.654)
			70.6	0.743 (0.558..0.928)	0.645 (0.465..0.826)	0.691 (0.563..0.819)

Wilcoxon rank-sum test	ADCk ICC>0.8 & Top 10	0.823(0.804..0.953) 0.823(0.804..0.953)	84.3	0.783 (0.637..0.929)	0.750 (0.592..0.908)	0.770 (0.664..0.875)

MRMR	ADCk ICC>0.8 & Top 10	0.888(0.808..0.962)
		0.836(0.800..0.962)
			74.5	0.727 (0.538..0.916)	0.732 (0.566..0.897)	0.730 (0.607..0.852)

Spearman ρ	ADCk ICC>0.8 & Top 10	0.904(0.800..0.951)
		0.927(0.800..0.942)
			72.5	0.686 (0.503..0.870)	0.723 (0.566..0.880)	0.706 (0.589..0.822)

AUC	ADCk ICC>0.8 & Top 10	0.904(0.800..0.951)
		0.927(0.800..0.942)
			72.5	0.686 (0.503..0.870)	0.723 (0.566..0.880)	0.706 (0.589..0.822)

Table 2:

Stability and performance of machine learning optimizing repeatability using average of feature values extracted from prostate DWI with four feature ranking methods. The 95% confidence intervals are shown in parenthesis. Rows 1–4: Agreement between classifications and performance for Gleason Grade Group (GGG) 1 vs >1 classification. Rows 5–8: Agreement between classifications and performance for GGG 1,2 vs >2.

Feature ranking method	**Pruned features**	**% of same labels**	**AUC (95% CI) Scan1**	**AUC (95% CI) Scan 2**	**AUC (95% CI) Scan 1 and 2**
Wilcoxon rank-sum test	ADCk & K ICC>0.8 & Top 10	60.8	0.780 (0.607..0.952)	0.757 (0.572..0.942)	0.765 (0.640..0.890)
MRMR	ADCk & K ICC>0.8 & Top 10	80.4	0.781 (0.601..0.960)	0.725 (0.560..0.890)	0.747 (0.629..0.866)
Spearman ρ	ADCk & K ICC>0.8 & Top 10	72.5	0.686 (0.503..0.870)	0.723 (0.566..0.880)	0.706 (0.589..0.822)
AUC	ADCk & K ICC>0.8 & Top 10	64.7	0.782 (0.601..0.962)	0.709 (0.548..0.870)	0.743 (0.623..0.863)
Wilcoxon rank-sum test	ADCk & K ICC>0.8 & Top 10	72.5	0.705 (0.530..0.879)	0.791 (0.600..0.981)	0.745 (0.620..0.870)
MRMR	ADCk & K ICC>0.8 & Top 10	66.7	0.734 (0.550..0.918)	0.745 (0.557..0.934)	0.732 (0.606..0.858)
Spearman ρ	ADCk & K ICC>0.8 & Top 10	62.7	0.745 (0.551..0.940)	0.734 (0.563..0.905)	0.738 (0.613..0.863)
AUC	ADCk & K ICC>0.8 & Top 10	62.7	0.745 (0.551..0.940)	0.734 (0.563..0.905)	0.738 (0.613..0.863)