Explainable artificial intelligence in prostate cancer treatment recommendation: A decision support system for oncological expert panels

Gregor Duwe, Dominique Mercier, Verena Kauth, Lisa Maria Jost, Kerstin Moench, Vikas Rajashekar, Markus Junker, Christopher C.M. Neumann, Andreas Dengel, Axel Haferkamp, Thomas Höfner

Supplementary Material

S1. Materials and Methods

S1.1. Ethic statement and legal framework

The study was conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. The data protection officer of University Medical Center Mainz (UMCM) and the ethics committee of the Medical Association of Rhineland-Palatinate, Germany (2022-16511, up-dated 2022-16511_2) approved the study. Finally, the legal cooperation (including data protection) between UMCM and the German Research Center for Artificial Intelligence (DFKI) was agreed by means of a cooperation agreement and an order processing agreement.

S1.2. Data security

Patient data sets were pseudonymised by the Institute for Medical Biostatistics, Epidemiology and Informatics of UMCM. A randomly generated 10-digit number was used as pseudonym. Personal data (name, date of birth, patient-ID) was removed before further processing and can only be restored by assigned staff members of the study team at UMCM for quality check of results.

S1.3. Data transfer between UMCM and German Research Center for Artificial Intelligence (DFKI)

To ensure a secure transfer, a locally hosted cloud storage at DFKI with an authentication service was used to upload the pseudonymised data by UMCM. Afterwards, the data was downloaded from the cloud storage and processed (preprocessed and trained) by DFKI. After the final evaluation using the AI-generated

treatment recommendation (TR) on the side of the storage, a result file is uploaded for the assessment by UMCM.

S1.4. Data preprocessing and network training

The downloaded data was technically adapted (“preprocessed”) to be suitable for AI development. As the tabular data contains complex patient characteristics, it was mandatory to simplify the complex data types without any loss of information. Numerical values were normalized and standardized to ensure optimal training of the machine learning architecture. Categorical values were additionally converted to numerical values. For the complex data types such as the programmed death-ligand 1 scores, splitting into different components was performed using regular expressions, followed by a categorization step. A two-step process was developed to train a classifier to mimic MCC recommendations. Due to the dataset size which was relatively small for a machine learning approach and the fact that multiple, equal treatment recommendations are given in some cases, a special processing step has been deployed (multi-label transformation via sample duplication) that makes the pipeline unique. Precisely, a methodological strategy (“counter intuitive decision making”) was established for MCCs with multiple recommendations which were duplicated with their different recommendations:

Since it was not possible to directly train a network for multi label prediction due to the low number of instances, a counter intuitive decision was established for MCCs with multiple equivalent recommendations by means of duplicating them to achieve single recommendations. It was ensured that such instances were always in the same (train or test) set. When a model is trained with duplicates, the distribution of prediction probabilities tends to be more evenly distributed rather than more predictive. A threshold can be used to determine the value at which a prediction is accepted as valid, making it feasible to specify multiple valid predictions with a single vector.

S2. Results – Supplementary information

S2.1. Confusion Matrices

The confusion matrix (Fig. S1) visualizes the overlaps between actual (true) and predicted TR, thereby showing how accurate the AI system's predictions are separately for each high-level and all low-level recommendations. Excellent overlap is

demonstrated for high-level recommendation anti-cancer drugs (Precision (P): 94%). The high-level TR surgery matches the true TR in most cases (P: 98%). Low-level recommendations for surgery, radiation therapy, radiopharmaceutical therapy and imaging show consistent concordance between the true and predicted TR. In the domain of low-level anti-cancer drug TR, an excellent overlap is achieved for ADT monotherapy (P: 94%). Overall, the lowest correspondence between predicted and true TR is observed here in the more variable field of combined drug therapies.

S2.2. Insertion scores

Insertion scores (Fig. S2) were generated complimentary to deletion scores (Fig. 4), to increase transparency on stability and robustness of the AI decision support system. They reveal the importance of specific parameters for decision-making, rated as most critical by the AI system. If decisions are based on irrelevant or distorted features, this may indicate a defective or biased decision-making process. The insertion scores facilitate a technical assessment of the content by the user, i.e. how sensible and comprehensible the decision-making strategy of the AI system is. The curve is plotted in a diagram displaying cross-section of F1-scores and number of patient related features changed. All insertion scores show a significantly increased curve progression and AUC and differ from random evaluation and classification of the patient related in- and output parameters. For both high- and low-level TR, they consistently yield significant results, indicating a comprehensible and reproduceable decision-making strategy of our AI system. The highest relative AUC increase is demonstrated for the low-level radiation recommendation.

Table S1. Clinical patient data structure:

76 Input Features (entry specified depending on patient data)

General patient data

- Age
- ECOG Performance Status

Comorbidities

- Arterial hypertension
- Cardiovascular diseases

- Renal insufficiency
- Degree of renal insufficiency
- Dialysis
- Neurological diseases
- Metabolic disorders
- Malignancies (other than MCC diagnosis)
- Treatment for further malignancies

Specific oncological data

- Initial diagnosis of prostate cancer – date
- Initial diagnosis of prostate cancer – date known or estimated (k/e)
- Diagnostic confirmation

Tumor markers

- Initial PSA value
- Current PSA value
- Nadir PSA value
- Time to PSA Nadir
- BRCA mutation

TNM Classification (histological)

- Histological type
- T-stage
- Lymphatic vessel invasion
- Venous invasion
- Perineural invasion
- N-stage
- M-stage
- Localization distant metastases
- Gleason Score
- Grading
- Risk classification (pretherapeutic)
- R-classification

- UICC

Previous anti-cancer drug treatments (for MCC diagnosis)

- Hormone-based therapy 1st line
- Hormone-based therapy 2nd line
- Hormone-based therapy 3rd line
- Hormone-based therapy 4th line
- Hormone-based therapy 5th line
- Hormone-based therapy 6th line
- Targeted cancer therapies
- Chemotherapy 1
- Chemotherapy 2
- Chemotherapy 3
- Radioligand therapy 1
- Radioligand therapy 2
- Supportive therapy

Previous anti-cancer treatment (other than drugs)

- Surgery
- Radiotherapy
- Other

Radiological imaging (most current staging, ≤3 months prior to MCC)

- Locoregional lymph node metastases
- Distant metastases
- Distant metastases localization

Laboratory values (most current analysis, ≤3 months prior to MCC)

- Leukocytes
- Hemoglobin
- Thrombocytes
- Neutrophil granulocytes
- GOT (ASAT)
- GPT (ALAT)
- gamma-GT
- Total bilirubin
- LDH

- Total protein
- Albumin
- Creatinine
- eGFR
- Sodium
- Potassium
- Calcium
- Urea-N
- Quick
- PSA

Current situation:

- Tumor progress
- Stable disease
- Tumor regress
- Mixed response
- Current UICC

MCC

- MCC date

23 Output Features (entry specified depending on patient data)

MCC recommendation

- Surgery
- Surgery (as additional equivalent recommendation)
- Radiotherapy
- Adjuvant Radiotherapy
- Anti-cancer drug treatment
- Anti-cancer drug treatment (as additional equivalent recommendation)
- Anti-cancer drug treatment (as additional equivalent recommendation 2)

- Anti-cancer drug treatment (as additional equivalent recommendation 3)
- Anti-cancer drug treatment (as additional equivalent recommendation 4)
- Supportive therapy
- PSMA ligand therapy
- PSMA ligand therapy (as additional equivalent recommendation)
- Active Surveillance
- Active Surveillance (as additional equivalent recommendation)
- Watchful Waiting
- Watchful Waiting (as additional equivalent recommendation)
- PSA Follow-up
- PSA Follow-up (as additional equivalent recommendation)
- Best supportive care
- Best supportive care (as additional equivalent recommendation)
- Imaging
- Prostate core biopsy
- Genetic diagnostics

Figures (S1 and S2)

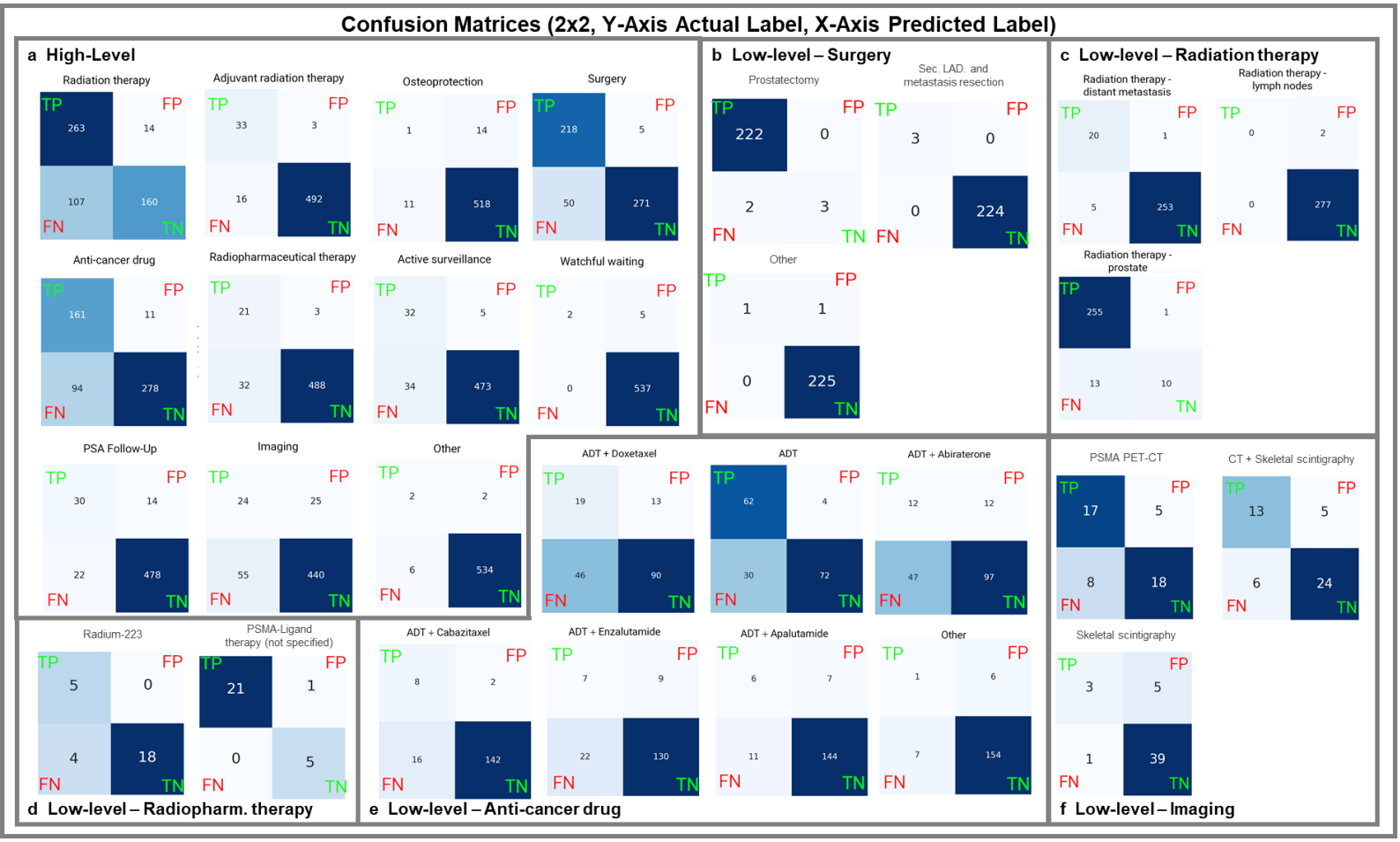

Fig. S1: Confusion Matrices to evaluate model performance for High-level TR (a) and Low-level TR (b-f). They summarize classification results by comparing the actual labels (MCC TR) (y-axis) with predicted labels of KITTU-XGB Classifier (x-axis). Each cell represents the number of samples assigned to a specific combination of actual and predicted labels. The darker the blue of a field, the more hits this combination receives. TP = true positive, FP = false positive, FN = false negative, TN = true negative, PSA=Prostate-Specific Antigen, Sec. LAD= secondary lymphadenectomy, ADT=Androgen-deprivation therapy, PSMA PET-CT= Prostate-Specific Membrane Antigen Positron Emission Tomography–Computed Tomography.

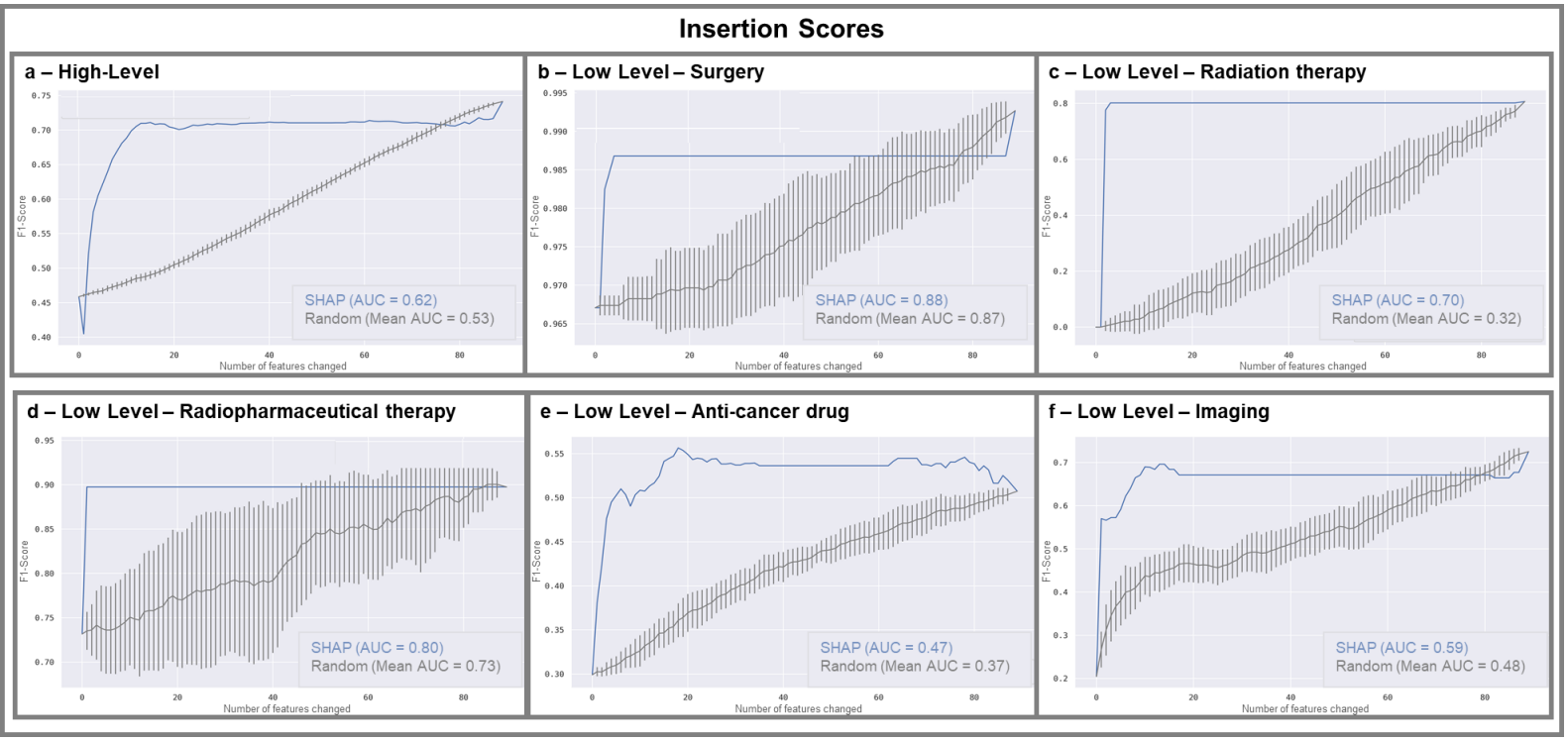

Fig. S2: Insertion scores for High-level (a) and Low-level (b-f) TR to evaluate relevance of features. Insertion score is displayed as a measure of explanatory quality. It is determined by gradually adding features to an initially meaningless baseline and then evaluating the model performance. SHapley Additive exPlanations (SHAP)-based assignments are compared with random mean baseline. Y-axis displays F1 Scores and x-axis displays number of features changed. Model performance is quantified using the area under the ROC curve (AUC); A faster increase in AUC indicates a more meaningful and reliable classification method.