International Journal of Medical Informatics 186 (2024) 105414

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

On relevant features for the recurrence prediction of urothelial carcinoma of the bladder

Louisa Schwarz [a,b,*], Dominik Sobania [b], Franz Rothlauf [b]

[a] Cancer Registry Rhineland-Palatinate, Mainz, Germany
[b] Johannes Gutenberg University, Mainz, Germany

ARTICLE INFO

Keywords:
Feature importance
Explainable artificial intelligence
XAI
Urothelial carcinoma of the bladder
Machine learning
Clinical decision support system

ABSTRACT

Background: Urothelial bladder cancer (UBC) is characterized by a high recurrence rate, which is predicted by scoring systems. However, recent studies show the superiority of Machine Learning (ML) models. Nevertheless, these ML approaches are rarely used in medical practice because most of them are black-box models, that cannot adequately explain how a prediction is made.
Objective: We investigate the global feature importance of different ML models. By providing information on the most relevant features, we can facilitate the use of ML in everyday medical practice.
Design, setting, and participants: The data is provided by the cancer registry Rhineland-Palatinate gGmbH, Germany. It consists of numerical and categorical features of 1,944 patients with UBC. We retrospectively predict 2-year recurrence through ML models using Support Vector Machine, Gradient Boosting, and Artificial Neural Network. We then determine the global feature importance using performance-based Permutation Feature Importance (PFI) and variance-based Feature Importance Ranking Measure (FIRM).
Results: We show reliable recurrence prediction of UBC with 82.02% to 83.89% F1-Score, 83.95% to 84.49% Precision, and an overall performance of 69.20% to 70.82% AUC on testing data, depending on the model. Gradient Boosting performs best among all black-box models with an average F1-Score (83.89%), AUC (70.82%), and Precision (83.95%). Furthermore, we show consistency across PFI and FIRM by identifying the same features as relevant across the different models. These features are exclusively therapeutic measures and are consistent with findings from both medical research and clinical trials.
Conclusions: We confirm the superiority of ML black-box models in predicting UBC recurrence compared to more traditional logistic regression. In addition, we present an approach that increases the explanatory power of black-box models by identifying the underlying influence of input features, thus facilitating the use of ML in clinical practice and therefore providing improved recurrence prediction through the application of black-box models.

1. Introduction

Around 430,000 people worldwide are diagnosed with bladder cancer every year. Most of them affect the urothelial of the bladder, a type of cancer which often recurs. Reliable prediction to determine recurrence of the tumor after successful cancer treatment is a relevant factor in treatment decisions [1]. Traditional approaches to calculate the recurrence risk of Urothelial Bladder Cancers (UBC) are EORTC [2] or CUETO [3]: Two scoring systems which classify the individual risk of recurrence in low, intermediate, and high-risk categories, which mostly determines further follow-up, such as frequency or the instillation of adjuvant chemotherapy [4,5].

Still, both systems have significant limitations which are discussed in [6–9]. First, they lack external validation verifying their predictive accuracy. Second, they are found to overestimate individual recurrence risk in high-risk patients when externally validated. And lastly, they do not include all relevant factors that might influence the development of bladder recurrence. Accordingly, each of these studies highlights the need for improved prognostic models.

Recent studies investigate machine learning (ML) models in the recurrence prediction of UBC, where the results are promising and show the potential to improve the prediction quality to consequently suggest a more appropriate and tailored follow-up [10,11]. However, ML approaches are rarely used as clinical decision support system (CDSS), as

* Corresponding author at: Johannes Gutenberg University, Mainz, Germany.
E-mail addresses: louisa.schwarz@uni-mainz.de (L. Schwarz), dsobania@uni-mainz.de (D. Sobania), rothlauf@uni-mainz.de (F. Rothlauf).

https://doi.org/10.1016/j.ijmedinf.2024.105414
Received 6 December 2023; Received in revised form 16 February 2024; Accepted 11 March 2024
Available online 18 March 2024

they still face some challenges in the prototyping environment, where further research is required [12–14].

One challenge is the black-box character of ML models, that cannot adequately explain how a prediction is made. As a result, medical professionals often distrust ML approaches, and mechanisms to increase explainability must be incorporated during the prototyping of such newer approaches [12,15,16]. An approach to increase explainability in ML is to evaluate the importance of input features on the given prediction within the entire model [17,18].

The goal of this work is to apply three supervised classification models with black-box character using Support Vector Machine [19], Gradient Boosting [20], and Artificial Neural Network [21] and a more traditional and explainable model using logistic regression as baseline to predict the 2-year recurrence of UBC. We then determine the retrospective global feature importance of the input features using two different model-agnostic methods and investigate whether those influencing features are consistent across the different models in the prediction of 2-year recurrence in patients with UBC.

The main contribution of this work is to show an approach which increases the explainability of black-box models with higher prediction quality by the investigation of the input feature importance on the prediction on a global scale. This approach could help medical professionals to better understand black-box models, thus enabling an increased use as CDSS, and therefore provide improved recurrence prediction by applying ML.

2. Related work

2.1. Recurrence prediction of urothelial carcinoma of the bladder

To ensure appropriate patient care under gold standards, medical guidelines with evidence-based statements are used in everyday clinical practice [1]. However, recent research indicates that the current gold standard for determining the risk of recurrence in UBC has limitations and improved prognostic models are required [6–9].

Further studies address this need for improved models in UBC recurrence prediction and develop appropriate models using ML. Table 1 gives an overview of related work which achieves reliable results for UBC recurrence prediction by ML models and summarizes them in terms of time-period, inclusion criteria, the data sets, and methods used as well as the results achieved.

All studies in Table 1 use data provided by different hospitals in different countries, published between 2009 and 2023. Three of the studies use tabular data [22,23,25], another studies supplement them with the extraction of information from image files, such as histopathological images [27–29] or magnetic resonance imaging (MRI) [26]. Frequently used features are age and sex of the patient, tumor size, ICD, morphology, and grading of the tumor. Some studies also include binary features on surgery or chemotherapy performed [22,26]. Furthermore, three of the studies include the lymph node status and lymphovascular invasion [25,22,23]. Another three studies exclusively examine patients with non-muscle invasive Bladder Cancer (NMIBC) [24,27,28], while some others analyze patients who have undergone transurethral resection (TUR) or radical cystectomy (RC) [23,24,27,29]. The observation period for recurrence ranges from six months to five years. Accordingly, the size of the data set varies from a few hundred to a few thousand patient data.

Despite their different data sets, all studies show reliable recurrence prediction in UBC. In addition, further studies indicate the superiority of those ML models over the traditional scoring systems [11].

2.2. Black-box characteristic

Despite the promising predictive quality of ML models (see Table 1), more traditional methods are primarily used for UBC recurrence prediction [1]. Related literature debates the use of ML models in high stakes decisions with black-box character, as it cannot adequately explain how a prediction is made or how the prediction is affected by the input features [30]. Further studies indicate that the black-box character of ML models is one of the factors preventing their use in clinical practice [15]. Accordingly, the use of ML in clinical practice faces some challenges and requires the incorporation of methods, which increase the explainability and provide insights into the prediction made within the prototyping process [30,12].

2.3. Explainable AI (XAI) in medicine

Explainable Artificial Intelligence (XAI) summarizes techniques that explain ML models and thus reduce their black-box character [17]. One method that increases the explainability of black-box models is to investigate the importance of input features and how they affect the given prediction on a global scale [17,18]. Additionally, the importance of each feature on the given prediction is relevant, especially when patient characteristics are not reflected in the input features of the prognostic model, such as secondary diseases, pregnancy, etc. [16]. To the best of our knowledge, this is the first study to investigate feature importance in the recurrence prediction of UBC.

3. Methodology

3.1. Dataset and patient selection

The data used is provided by the German epidemiological and clinical Cancer Registry Rhineland-Palatinate, which collects and processes data on the diagnosis, therapy, and follow-up of tumor patients within Rhineland-Palatinate, which are reported by all treating physicians. All patients are informed about the processing of their data in accordance with the federal law of Cancer Registration. The retrospective analysis includes patients diagnosed with UBC between 01/01/2016 and 12/31/2019, including treatment and progression until 12/31/2021. Only male or female patients older than 18 were included as well as patients who underwent TUR and/or RC. According to these criteria, we observe a total of 1,944 patients with an average age of 70.4 years ($\pm$ 10.6), of which 79.58% are male and 20.42% female.

3.2. Data cleansing and feature selection

All data is available in raw tabular form, whereby each data record refers to one patient. The selection of input features follows the features used in clinical practice [5,4], features used in related work (see Table 1), as well as further time-bound and therapeutic features. Information on comorbidities, ethnic group or socioeconomic status is not available. Table 2 summarizes the input features.

Data cleansing is performed carefully in collaboration with cancer registry medical specialists to prevent serious malfunction and ensure high quality data [31,32]. For the features of TNM classification, GRADING, and ICD, the most severe expression per patient is taken into account. Given the TNM-features (T, N, M, L, V) include the valid category *X* for *No specification possible* in their respective classifications, all missing values are coded to their respective *X* before preprocessing.

3.3. Experimental setting and preprocessing

This study focuses on retrospective 2-year recurrence prediction of UBC after completed treatment of the primary tumor [1]. Among the 1,944 patients studied, recurrence occurs in 21.14% of patients with corresponding 78.86% of patients surviving free of recurrence. We perform a binary classification with *recurrence* as positive and *no recurrence* as negative class. All analyses are performed using R in the Tidymodels framework.

The data is first split into 80% training and 20% testing data. Using training data, the prediction model $\hat{f}$ learns the relation between

Table 1
Overview of related work comparing data, methods, and results for recurrence prediction in UBC by ML.

Ref.	Year	Period	Inclusion criteria	Data set and features	#patients	Methods	Results on testing data				
							F1-Score	Sens	Spec	Prec	AUC
[22]	2009	5-year	All UBC (NMIBC and MIBC)	Clinical and pathological information from tabular data	609	ANN	X	0.81	0.85	X	X
[23]	2013	5-year	All UBC after TUR or RC	Clinical and pathological information from tabular data	2,111	ANN	X	0.4	0.9	0.63	X
[24]	2016	5-year	Only NMIBC after TUR	Genes from genome profiling from frozen specimens	112	GP	X	0.71	0.67	X	X
[25]	2019	1-year 3-year 5-year	All UBC (NMIBC and MIBC)	Clinical and pathological information from tabular data	3,071 2,955 2,695	Meta-Classifier	0.51 0.61 0.64	0.74 0.78 0.70	0.71 0.71 0.70	0.39 0.54 0.59	X X X
[26]	2019	2-year	All UBC (NMIBC and MIBC)	Radiomics and clinical information from MRI images, and tabular data	71	Nomogram	0.81	X	X	X	0.84
[27]	2021	2-year	Only NMIBC after TUR	Clinical and pathological information from histological images, and tabular data	125	SVM RF	0.83 0.75	0.88 0.80	0.80 0.70	0.79 0.69	X X
[28]	2022	1-year	Only NMIBC	Clinical and pathological information from histopathological images, and tabular data	359	Network-based models	X	0.3	0.8	0.44	0.62
		5-year			281		X	0.89	0.57	0.71	0.76
[29]	2023	6-mos to 5-year	All UBC after TUR or RC	Clinical information from CT-images, and tabular data	874	Network-based models	X	X	X	X	0.89

UBC = Urothelial bladder cancer; NMIBC = Non-muscle invasive bladder cancer; MIBC = Muscle-invasive bladder cancer; TUR = Transurethral resection, RC = Radical cystectomy; MRI = Magnetic Resonance Imaging; ANN = Artificial neural network; GP = Genetic programming; SVM = Support vector machine; RF = Random forest; Sens = Sensitivity; Spec = Specificity; Prec = Precision or Positive Predictive Value; AUC = Area under ROC-Curve.

the input feature matrix X and the output feature vector y, attempting to minimize the training classification error e. For training data, we perform the following pre-processing steps: First, missing values in two features are imputed according to the specifications in Table 3. Second, all categorical features are encoded into numerical ones, since not all ML algorithms accept categorical data types. All ordered categorical features are ordinal encoded regarding the underlying order while all unordered categorical features are one-hot encoded. Third, all data is normalized with a mean per feature of zero and a standard deviation of one to generate a uniform scale for all input features. And lastly, we apply stratified Borderline SMOTE [33] with respect to the target feature as sampling technique to counteract the problem of an imbalanced dataset, since ML models tend to always predict the overrepresented class (see Sect. 3.1).

After training, the testing data is used to evaluate the ML model. To ensure a realistic evaluation of the models, the testing data is not sampled. Based on the correctly and incorrectly classified labels in the testing data, we calculate the main performance metrics for each model, which are presented in more detail in Sect. 4.1.

3.4. Feature importance for machine learning models

To study the influence of input features on recurrence prediction in UBC, we first apply three black-box models namely Artificial feed-forward Neural Networks (ANN), Gradient Boosting (GB), and Support Vector Machine (SVM) as representative models of the methods used in Table 1 and compare their results to Logistic Regression (LR) as baseline. Appendix A provides a detailed overview of the parameter settings per model.

Second, we use two different model-agnostic measurement methods to investigate the underlying feature importance of each input feature vector $x_i \in x_1, ..., x_n$ in an input feature matrix X to predict the target feature vector y for a trained model $\hat{f}$.

First, Permutation Feature Importance (PFI). Initially, we train $\hat{f}$ given X to predict y and compute the default resulting error $e_{\text{origin}} = L(y, \hat{f}(X))$, where $L()$ is the used loss function. Second, each feature x_i in X is consecutively permuted, which disconnects the relationship between x_i and y. We perform a new prediction y for the permutated input and compute the permutation error $e_{i,\text{perm}} = L(y, \hat{f}(x_{i,\text{perm}}, X_{\setminus i}))$ given X with x_i replaced by $x_{i,\text{perm}}$. The resulting PFI is calculated as

$$\text{PFI}_i = e_{i,\text{perm}} - e_{\text{origin}}. \tag{1}$$

Higher errors $e_{\text{i,perm}}$ lead to higher PFI_i and to a higher importance of feature x_i for the prediction quality of $\hat{f}$. In this work, the loss is calculated as $L = 1 - \text{AUC}$. For more detail, see [34,35].

Second, Feature Importance Ranking Measure (FIRM) is a variance-based approach, introduced by [36]. It retrospectively analyses the importance of x_i in X for the prediction of y in $\hat{f}$ in a two step-approach. First, the so called estimated conditional expected score (CES) for each feature x_i in X is defined as $\text{CES}_i = \hat{f}(x_i, X_{\setminus i})$, whereby it estimates the partial dependence of each feature x_i on the trained model $\hat{f}$ given X. Second, the FIRM for x_i is calculated as

$$\text{FIRM}_i = \begin{cases} \sqrt{\text{Var}(\text{CES}_i)} & x_i = \text{numerical} \\ \frac{1}{4}(\max(\text{CES}_i) - \min(\text{CES}_i)) & x_i = \text{categorical.} \end{cases} \tag{2}$$

Higher values of CES_i lead to higher FIRM_i and to a higher importance of feature x_i for $\hat{f}$ [37].

4. Results

4.1. Model performance

To study the feature importance of black-box models, we first apply different ML models. We use Accuracy, F1-Score, Precision, and Area Under the curve (AUC) as performance metrics, whereby *recurrence* is labeled as positive and *no recurrence* as negative class. In particular, the F1-Score is an appropriate metric because it takes into account the prediction of the positive class and is suitable for imbalanced data sets [38]. F1-Score is the harmonic mean between precision and sensitivity, where

Table 2
Patient, tumor, and therapy characteristics of the data set provided by the Cancer Registry Rhineland-Palatinate, separately for categorical and numerical features.

Numeric features

Name	*Unit*	*Range*	*Mean*	*Std. Dev.*
AGE	years	31-97	70.4	± 10.6
OP_TUR	count	0-7	1.49	± 0.867
SYS_CH	count	0-6	0.202	± 0.578
SYS_IM	count	0-7	0.069	± 0.368
N_DAYS	days	0-1745	25.1	± 116.0

Categorical features

Name	*Ordered*	*Codes*	*Count*	*Distribution*
SEX	false	male	1547	0.7958
		female	397	0.2042
ICD	false	D41.4 + D09.0	988	0.5082
		C67	956	0.4918
MORPH	false	8130/1 + 8120/2 + 8130/2	1464	0.7531
		8120/3	480	0.2469
TNM_T	true	TX	60	0.0309
		Tis	41	0.0211
		Ta	933	0.4799
		T1	544	0.2798
		T2	244	0.1255
		T3	100	0.0514
		T4	22	0.0113
TNM_N	true	NX	774	0.3981
		N0	1101	0.5664
		N1	36	0.0185
		N2	22	0.0113
		N3	11	0.0057
TNM_M	true	MX	624	0.3210
		M0	1304	0.6708
		M1	16	0.0082
TNM_L	true	LX	1482	0.7623
		L0	349	0.1795
		L1	113	0.0581
TNM_V	true	VX	1495	0.7690
		V0	398	0.2047
		V1	51	0.0262
GRADING	true	U	78	0.0401
		L	985	0.5067
		M	7	0.0036
		H	874	0.4496
OP_RC	false	yes	307	0.1579
		no	1637	0.8421

ICD = ICD-Code; MORPH = Morphology; TNM_T = tumor stage; TNM_N = lymph node stage; TNM_M = metastasis stage; TNM_L = lympathic vessels invasion stage; TNM_V = vein invasion stage; OP_RC = performance of RC; OP_TUR = number of TUR performed; SYS_CH = performance of chemotherapy; SYS_IM = performance of hormonetherapy; N_DAYS = number of days between initial diagnosis and initial treatment.

sensitivity indicates the proportion of the correctly predicted positive class to the true positive class. Precision indicates the proportion of the correctly predicted positive class to the whole predicted positive class.

Table 4 shows performance metrics on testing and training data, which reflect the average and standard deviation of 30 runs. All three black-box models – ANN, GB, and SVM – significantly outperform LR in Accuracy and F1-Score. In addition, the three models achieve comparable results to those in related work (see Table 1). The black-box models

Table 3
Overview of features containing missing values.

Name	Count	Percentage	Imputation
GRADING	78	0.0401	Mode
N_DAYS	102	0.0525	Mean

GRADING: Missing values are assigned to the most frequently represented category. N_DAYS: This feature is calculated from the date of diagnosis and the date of the first treatment. In 102 patient, the calculation results in a negative and therefore incorrect and missing value, with *mean* imputation applied.

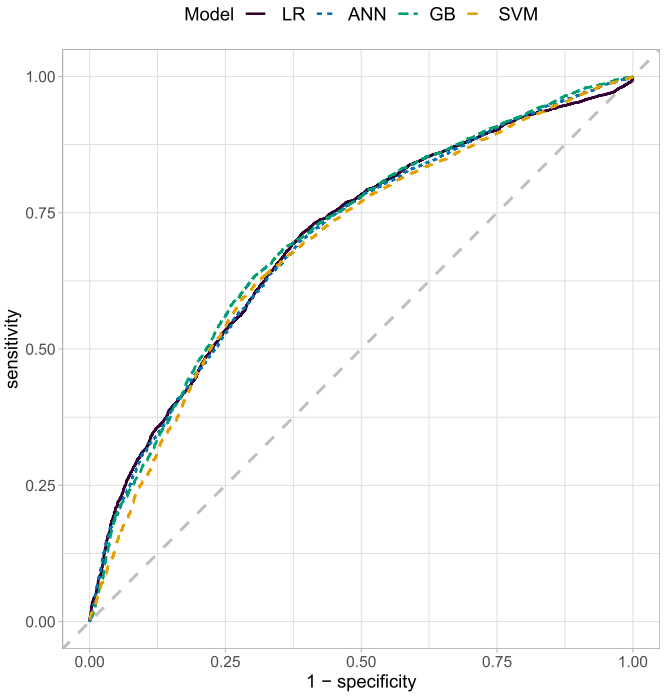

Fig. 1. ROC-curve for recurrence prediction in UBC across all models (ANN, GB, SVM, LR).

correctly predict recurrence in UBC between 72.51% and 74.68%, while LR performs significantly worse at 70.84%. In addition, the prediction of the positive class is found to be appropriate and significantly better than LR, with F1-Score ranging from 82.02% to 83.89% depending on the model. The small differences between the testing and training results indicate generalization of all models, resulting in reliable predictions for previously unseen data. Accordingly, the models do not tend to overfit, indicating sufficient sample size [39].

The predictive performance is represented by the receiver operating characteristic (ROC) curve, shown in Fig. 1. The ROC-curve of each ML model shows the respective trade-off between sensitivity and specificity using their calculated probabilities. Thus, the closer the ROC-curve moves to the y-axis and the farther away it moves from the x-axis, the better the model performs in predicting correctly, while the gray dashed line shows the performance of a random model. All ROC-curves are clearly different from the performance of a random model, although the black-box models do not perform significantly better than LR. GB shows the best Area Under the Curve (AUC) of 70.82%, closely followed by the ANN und LR with 70.22% and 70.21% respectively, the SVM shows a slightly lower AUC of 69.20% (see Table 4).

The ROC-curves do not differ substantially, and the models predict all regions of the curves similarly. In a medical context, however, the interpretation of ROC-curves is often too general and does not provide a deeper analysis for individual groups, such as high-risk patients [40].

Table 4
Performance of the 2-yr-recurrence prediction for UBC by different models on testing and training data. Results reflect the mean and standard deviation of 30 runs for each model, with the best mean per metric and model in bold. Wilcoxon rank-sum tests with Holm correction are performed on the testing results with LR as reference group. Statistical significance ($p < 0.01$) of each model compared to LR is indicated by *.

Testing data

Metric	ANN	GB	SVM	LR
Accuracy	0.7320* ± 0.0277	**0.7468*** ± 0.0171	0.7251* ± 0.0222	0.7084 ± 0.0213
F1-Score	0.8255* ± 0.0223	**0.8389*** ± 0.0131	0.8202* ± 0.0182	0.8019 ± 0.0167
Precision	0.8449 ± 0.0150	0.8395 ± 0.0120	0.8442 ± 0.0120	**0.8614** ± 0.0139
AUC	0.7022 ± 0.0298	**0.7082** ± 0.0245	0.6920 ± 0.0285	0.7021 ± 0.0289

Training data

Metric	ANN	GB	SVM	LR
Accuracy	0.7558 ± 0.0170	**0.7912** ± 0.0125	0.7447 ± 0.0126	0.7132 ± 0.0085
F1-Score	0.8418 ± 0.0149	**0.8678** ± 0.0103	0.8342 ± 0.0107	0.8058 ± 0.0074
Precision	0.8597 ± 0.0100	**0.8660** ± 0.0077	0.8550 ± 0.0066	0.8650 ± 0.0060
AUC	0.7305 ± 0.0111	**0.7844** ± 0.0095	0.7303 ± 0.0113	0.7126 ± 0.0095

F1-Score = harmonic mean between precision and sensitivity (true positive rate).
Precision = positive predictive value; AUC = Area under the curve.

Appendix B shows calibration curves that provide a detailed analysis of predicted probabilities for individual groups.

4.2. Feature importance measurement

Figs. 2 and 3 show the results of feature importance across all studied ML models, calculated by PFI and FIRM. The input features are arranged on the y-axis from top to bottom in descending order of average importance for all three models, while the x-axis shows the calculated importance.

For PFI (Fig. 2), the importance of each input feature is mostly consistent across all models. Most of the features are ranked in the same order across the models, in particular for the first four most important features, which represent therapeutic measures only, i.e., the number of TUR performed (OP_TUR), the performance of RC (OP_RC), the number of days between diagnosis and therapy (N_DAYS), and any additional chemotherapy administered (SYS_CH). These results indicate that all models identify the same relations between input and output features as the most important ones and accordingly apply these insights on the recurrence prediction in UBC.

For FIRM (Fig. 3), we observe similar results. Again, the top four features are consistent with those on PFI, only the ranked order differs slightly. GB and SVM agree on the order of important input features, also ranking OP_TUR as most important. Compared to PFI, N_DAYS is ranked second, followed by OP_RC and SYS_CH. For LR, the order is slightly different, with SYS_CH considered the most important feature, followed by therapeutic features OP_TUR, SYS_IM, and N_DAYS. Only for ANN, TNM_V is ranked among the four most important features (in 5th place on average on PFI), in addition to the features OP_RC, OP_TUR and SYS_CH. Therefore, FIRM also identifies the therapeutic features to be most relevant, meaning these results are consistent with PFI.

These results are consistent with related literature and confirm previous findings that treatment factors have a significant impact on recurrence in UBC. Several studies agree on the impact of surgical decisions such as OP_TUR and OP_RC [23,26,28]; further studies improve the prognostic quality by including information on systemic chemotherapy (SYS_CH) [22,27,28]. In addition, clinical evidence is provided on the impact of choice and timing of therapy on recurrence in UBC. Clinical studies show that starting therapy after three months has a significantly negative effect, making N_DAYS an important factor in UBC recurrence prognosis [41].

In conclusion, our results confirm that future prognostic tools for recurrence in UBC should include all available information on therapy. Our results also show that they are consistent with findings from related work and clinical trials, where the same therapeutic features are identified as important. Consequently, both feature importance methods provide insight into how the ML models generate their predictions on a global scale, thereby reducing their black-box character.

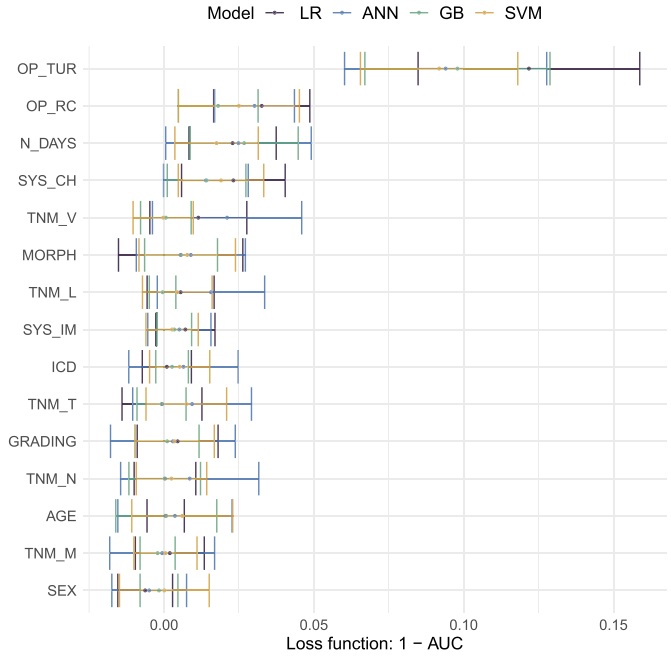

Fig. 2. Performance-based Permutation Feature Importance (PFI) of all features on recurrence prediction in UBC across all models (ANN, GB, SVM, LR).

5. Discussion

5.1. Conclusion

This study aims to increase the explainability of ML models for the recurrence prediction in UBC. Most of ML models are black-box approaches, which prevent their use in practice as CDSS, although their performance is more reliable compared to currently used systems [11]. Therefore, we studied the influence of input features on the recurrence prediction of UBC, while first applying three different ML models vs. a more traditional and explainable model using Logistic Regression and second measuring the importance of input features using two different feature importance measurement methods.

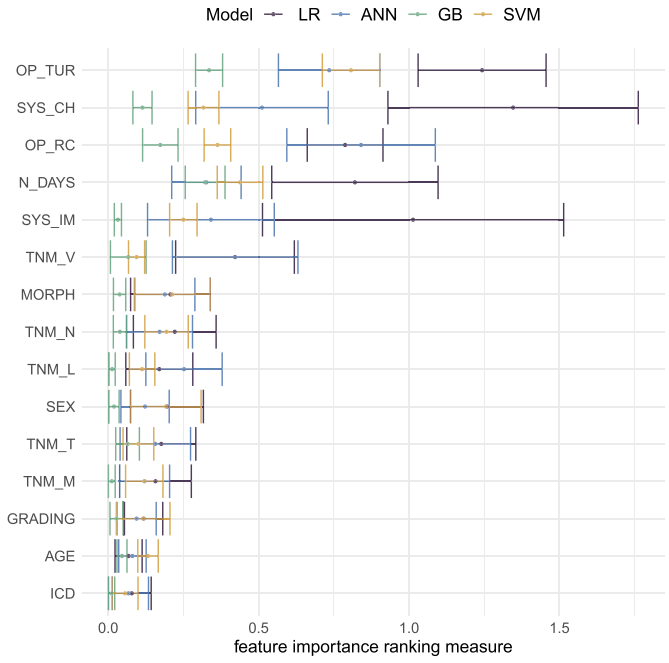

Fig. 3. Variance-based Feature Importance Ranking Measure (FIRM) of all features on recurrence prediction in UBC across all models (ANN, GB, SVM, LR).

Measured by F1-Score and accuracy, all black-box models performed significantly better compared to the LR. Furthermore, we identified predominantly the same features as most important ones, which are exclusively therapeutic measures, namely the number of TUR performed, the performance of RC, the number of days between initial diagnosis and initiation of therapy, and the administration of chemotherapy, which are consistent with the medical research [1,23]. In particular, therapeutic interventions have an impact on recurrence prediction in UBC, as shown by all black-box models in this paper and confirmed by previous results in related work using ML models (see Table 1) as well as clinical studies [41]. In conclusion, all black-box models demonstrate reliable prediction and provide important insights into features that should be considered in the future when predicting recurrence in UBC.

We recommend including both feature importance measurement methods in practice to provide both consistent and clear results for identifying important input factors, as the permutation results are more consistent, while the FIRM sometimes allows clearer statements.

We presented an approach that can increase the explainability of black-box models by identifying the underlying influence of the input features on the given prediction. Accordingly, the transparency of black-box models is increased allowing medical professionals to better understand the prediction. In conclusion, the investigation of feature importance facilitates the use of ML models by addressing the lack-of-transparency problem.

5.2. Limitations and future work

To achieve better prediction results, hyper-parameter optimization will be performed for all black-box models in future studies. In addition, other approaches to XAI will be explored, as this study focuses exclusively on feature importance on a global scale. Furthermore, these models could be tested for external validation in clinical studies.

6. Summary table

What was already known on the topic:

- Currently used scoring systems (EORTC and CUETO) lack external validation and are limited in their performance compared to newer machine learning (ML) models.
- ML models are rarely used as clinical decision support system (CDSS), as most of them do not provide transparency about how a prediction is made.

What this study added to our knowledge:

- We presented an approach that can increase the explainability of black-box models by identifying the underlying influence of input features.
- Investigating the importance of features allows medical professionals to better understand the prediction and facilitates the use of ML models in medical practice by addressing the lack of transparency problem.
- When predicting the recurrence of UBC using ML models, the most relevant features are the performance of transurethral resections or radical cystectomy, the administration of chemotherapy, and the number of days between diagnosis and therapy. These features are consistent with medical research.

CRediT authorship contribution statement

Louisa Schwarz: Writing – original draft, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dominik Sobania:** Writing – review & editing, Visualization, Supervision, Conceptualization. **Franz Rothlauf:** Writing – review & editing, Supervision, Resources, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interests to declare. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used DeepL Translate and DeepL Write in order to check spelling and grammar. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgement

The research has been declared as ethically unobjectionable according to the Joint Ethics Committee of the Faculty of Economics and Business Administration of Goethe University Frankfurt and the Gutenberg School of Management & Economics of the Faculty of Law, Management and Economics of Johannes Gutenberg University Mainz. The authors thank the Cancer Registry Rhineland-Palatinate for providing the patient data studied in this paper.

Appendix A. Machine learning models

To study the influence of input features on recurrence prediction in UBC, we first apply three black-box ML models as representative models of the methods used in Table 1 and compare their results to Logistic Regression (LR) as baseline:

- Artificial feed-forward Neural Networks (ANN) consist of neurons arranged in layers, which pass on the information of the input features via the internal hidden layers by using non-linear activation functions. The respective weights and consequently the information passing from one neuron to another is adjusted during the training process to minimize the resulting classification error [21].

Table 5
Parameter settings for used ML black-box models for binary classification.

Model	Parameter	Setting
Gradient Boosting	n_trees	15
	learn_rate	0.1
	tree_depth	6
SVM	cost	0.1
	margin	0.1
ANN	epochs	100
	penalty	0.1
	hidden_units	2
	dropout	0
	activation	ReLU

- Gradient Boosting (GB) models are decision tree-based ensembles. They consist of uncorrelated decision trees which are generated sequentially from bootstrap samples of the training dataset, existing trees are not updated. The resulting error is calculated using a loss function. When generating further decision trees within the ensemble, parameters are iteratively adjusted according to the gradient descent procedure to minimize the error calculated by the loss function [20].
- Support Vector Machines (SVM) separate two classes from each other by selecting a hyperplane in a n-dimensional space where the distance of the hyperplane to each class is maximized (maximum-margin hyperplane). A soft-margin defines how many misclassifications are allowed when applying this maximum-margin hyperplane. If the most suitable solution is a non-linear separation of both classes, additional kernel functions allow the data to be projected to a higher dimensional space, thus achieving a non-linear separation of both classes [19]. A radial basis function kernel parameter is used to maximize the distance between the two classes.
- LR uses regression coefficients for its input features, and is based on the maximum likelihood estimation. LR is considered a traditional model and has intrinsic transparency due to the coefficients that imply the importance of input features [42].

The parameter setting for each model is presented in Table 5, further hyper-parameter optimization is not performed.

Appendix B. Calibration

Medical prediction models require a calibration so that the predicted probability-like scores reflect the distribution of the true classes. To recalibrate the predicted probabilities to the true distribution of the classes, post-processing models are trained. In this work, we use a sigmodial logistic regression model and a Naïve Bayes approach [43].

For all models, the predicted probabilities are grouped into 10 bins and compared to the true distribution of the classes. Fig. 4 plots calibration curves for the two models; the gray dashed line indicates perfect calibration. The models are found to be overconfident in predicting the positive class, mostly due to the imbalanced data set and the small number of positive class cases, resulting in a biased representation. Nevertheless, all three models show a highly accurate calibration when predicting the negative class, as indicated by the calibration curves near the gray dashed line.

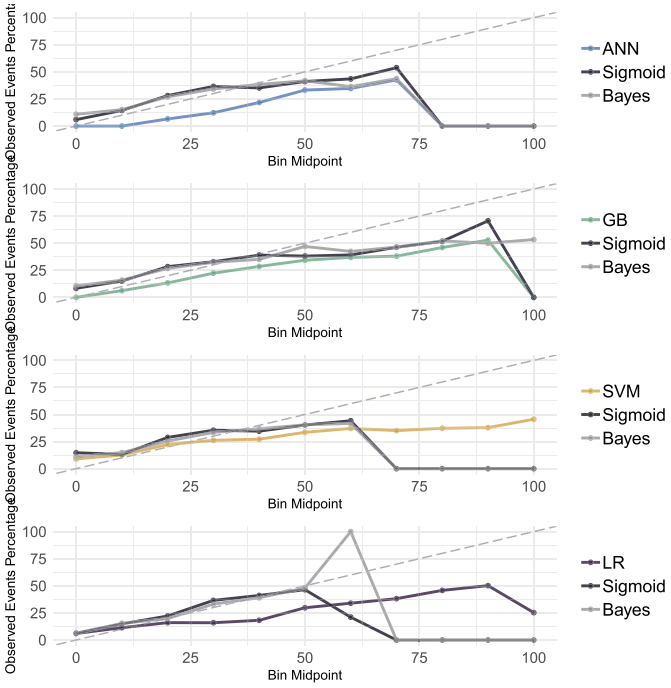

Fig. 4. Calibration curves across the different models for ANN (top), GB, SVM, and LR (bottom). Results for the sigmodial logistic regression model and naive Bayes approach are ploted by black and grey lines, respectively.

References

[1] A.M. Kamat, N.M. Hahn, J.A. Efstathiou, S.P. Lerner, P.-U. Malmström, W. Choi, C.C. Guo, Y. Lotan, W. Kassouf, Bladder cancer, Lancet 388 (2016) 2796–2810.
[2] R.J. Sylvester, A.P. Van Der Meijden, W. Oosterlinck, J.A. Witjes, C. Bouffioux, L. Denis, D.W. Newling, K. Kurth, Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using eortc risk tables: a combined analysis of 2596 patients from seven eortc trials, Eur. Urol. 49 (2006) 466–477.
[3] J. Fernandez-Gomez, R. Madero, E. Solsona, M. Unda, L. Martinez-Piñeiro, M. Gonzalez, J. Portillo, A. Ojea, C. Pertusa, J. Rodriguez-Molina, et al., Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the cueto scoring model, J. Urol. 182 (2009) 2195–2203.
[4] M. Babjuk, W. Oosterlinck, R. Sylvester, E. Kaasinen, A. Böhle, J. Palou-Redorta, Eau guidelines on non-muscle-invasive urothelial carcinoma of the bladder, Eur. Urol. 54 (2008) 303–314.
[5] J.A. Witjes, T. Lebret, E.M. Compérat, N.C. Cowan, M. De Santis, H.M. Bruins, V. Hernández, E.L. Espinós, J. Dunn, M. Rouanne, et al., Updated 2016 eau guidelines on muscle-invasive and metastatic bladder cancer, Eur. Urol. 71 (2017) 462–475.
[6] M. Jobczyk, K. Stawiski, W. Fendler, W. Różański, Validation of eortc, cueto, and eau risk stratification in prediction of recurrence, progression, and death of patients with initially non–muscle-invasive bladder cancer (nmibc): a cohort analysis, Cancer Med. 9 (2020) 4014–4025.
[7] A. Dalkilic, G. Bayar, M.F. Kilinc, A comparison of eortc and cueto risk tables in terms of the prediction of recurrence and progression in all non-muscle-invasive bladder cancer patients, Urol. J. 16 (2019) 37–43.
[8] M.C. Leo, C.K. McMullen, M. O'Keeffe-Rosetti, S. Weinmann, T. Garg, M.E. Nielsen, External Validation of the Eortc and Nccn Bladder Cancer Recurrence and Progression Risk Calculators in a US Community-Based Health System, Urologic Oncology: Seminars and Original Investigations, vol. 38, Elsevier, 2020, pp. 39–e21.
[9] E. Xylinas, M. Kent, L. Kluth, A. Pycha, E. Comploj, R. Svatek, Y. Lotan, Q. Trinh, P. Karakiewicz, S. Holmang, et al., Accuracy of the eortc risk tables and of the cueto scoring model to predict outcomes in non-muscle-invasive urothelial carcinoma of the bladder, Br. J. Cancer 109 (2013) 1460–1466.
[10] R. Suarez-Ibarrola, S. Hein, G. Reis, C. Gratzke, A. Miernik, Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer, World J. Urol. 38 (2020) 2329–2347, https://doi.org/10.1007/s00345-019-03000-5.
[11] C. Gandi, L. Vaccarella, R. Bientinesi, M. Racioppi, F. Pierconti, E. Sacco, Bladder cancer in the time of machine learning: intelligent tools for diagnosis and management, Urol. J. 88 (2021) 94–102, https://doi.org/10.1177/0391560320987169.
[12] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (2017) 230–243, https://doi.org/10.1136/svn-2017-000101.
[13] S. Borhani, R. Borhani, A. Kajdacsy-Balla, Artificial intelligence: a promising frontier in bladder cancer diagnosis and outcome prediction, Crit. Rev. Oncol./Hematol. 171 (2022), https://doi.org/10.1016/j.critrevonc.2022.103601.
[14] D. van de Sande, M.E. van Genderen, J. Huiskens, D. Gommers, J. van Bommel, Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit, Intensive Care Med. 47 (2021) 750–760, https://doi.org/10.1007/s00134-021-06446-7.

[15] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review, Appl. Sci. 11 (2021) 5088.
[16] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J.H. Chen, X. Liu, Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inform. Assoc. 27 (2020) 1173–1185, https://doi.org/10.1093/jamia/ocaa053.
[17] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[18] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018), https://doi.org/10.1145/3236009.
[19] W.S. Noble, What is a support vector machine?, Nat. Biotechnol. 24 (2006) 1565–1567.
[20] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232.
[21] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[22] J.W. Catto, M.F. Abbod, D.A. Linkens, S. Larré, D.J. Rosario, F.C. Hamdy, Neurofuzzy modeling to determine recurrence risk following radical cystectomy for nonmetastatic urothelial carcinoma of the bladder, Clin. Cancer Res. 15 (2009) 3150–3155.
[23] A. Buchner, M. May, M. Burger, C. Bolenz, E. Herrmann, H.-M. Fritsche, J. Ellinger, T. Höfner, P. Nuhn, C. Gratzke, et al., Prediction of outcome in patients with urothelial carcinoma of the bladder following radical cystectomy using artificial neural networks, Eur. J. Surg. Oncol. 39 (2013) 372–379.
[24] G. Bartsch Jr, A.P. Mitra, S.A. Mitra, A.A. Almal, K.E. Steven, D.G. Skinner, D.W. Fry, P.F. Lenehan, W.P. Worzel, R.J. Cote, Use of artificial intelligence and machine learning algorithms with gene expression profiling to predict recurrent nonmuscle invasive urothelial carcinoma of the bladder, J. Urol. 195 (2016) 493–498.
[25] Z. Hasnain, J. Mason, K. Gill, G. Miranda, I.S. Gill, P. Kuhn, P.K. Newton, Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients, PLoS ONE 14 (2019) e0210976.
[26] X. Xu, H. Wang, P. Du, F. Zhang, S. Li, Z. Zhang, J. Yuan, Z. Liang, X. Zhang, Y. Guo, et al., A predictive nomogram for individualized recurrence stratification of bladder cancer using multiparametric mri and clinical risk factors, J. Magn. Reson. Imaging 50 (2019) 1893–1904.
[27] N. Tokuyama, A. Saito, R. Muraoka, S. Matsubara, T. Hashimoto, N. Satake, J. Matsubayashi, T. Nagao, A.H. Mirza, H.-P. Graf, et al., Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features, Mod. Pathol. 35 (2022) 533–538.
[28] M. Lucas, I. Jansen, T.G. van Leeuwen, J.R. Oddens, D.M. de Bruin, H.A. Marquering, Deep learning–based recurrence prediction in patients with non–muscle-invasive bladder cancer, Eur. Urol. Focus (2020).
[29] H. Wang, M. Zhang, J. Miao, F. Hou, Y. Chen, Y. Huang, L. Yang, S. Yang, C. Huang, Y. Song, et al., Deep learning signature based on multiphase enhanced ct for bladder cancer recurrence prediction: a multi-center study, eClinicalMedicine 66 (2023).
[30] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.
[31] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: the European legal perspective, Commun. ACM 64 (2021) 34–36.
[32] K. Stöger, D. Schneeberger, P. Kieseberg, A. Holzinger, Legal aspects of data cleansing in medical AI, Comput. Law Secur. Rev. 42 (2021) 105587.
[33] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International Conference on Intelligent Computing, Springer, 2005, pp. 878–887.
[34] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[35] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 1–81.
[36] A. Zien, N. Krämer, S. Sonnenburg, G. Rätsch, The feature importance ranking measure, 2009, pp. 694–709.
[37] C.A. Scholbeck, C. Molnar, C. Heumann, B. Bischl, G. Casalicchio, Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations, Springer International Publishing, 2020.
[38] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (2022) 5979.
[39] X. Ying, An overview of overfitting and its solutions, J. Phys. Conf. Ser. 1168 (2019) 022022, IOP Publishing.
[40] A.M. Carrington, D.G. Manuel, P.W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, et al., Deep roc analysis and auc as balanced average accuracy, for improved classifier selection, audit and explanation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 329–341.
[41] N.M. Fahmy, S. Mahmud, A.G. Aprikian, Delay in the surgical treatment of bladder cancer and survival: systematic review of the literature, Eur. Urol. 50 (2006) 1176–1182.
[42] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, vol. 398, John Wiley & Sons, 2013.
[43] M. Kuhn, K. Johnson, Applied Predictive Modeling, vol. 26, Springer, 2013.