International Journal of
Molecular Sciences

MDPI

Article

Computational Analysis Identifies Novel Biomarkers for High-Risk Bladder Cancer Patients

Radosław Piliszek [1,*], Anna A. Brożyna [2] and Witold R. Rudnicki [1,3,*]

[1] Computational Centre, University of Białystok, ul. Konstantego Ciołkowskiego 1M, 15-245 Białystok, Poland
[2] Department of Human Biology, Institute of Biology, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University, ul. Lwowska 1, 87-100 Toruń, Poland; anna.brozyna@umk.pl
[3] Institute of Computer Science, University of Białystok, ul. Konstantego Ciołkowskiego 1M, 15-245 Białystok, Poland
* Correspondence: r.piliszek@uwb.edu.pl (R.P.); w.rudnicki@uwb.edu.pl (W.R.R.)

Abstract: In the case of bladder cancer, carcinoma in situ (CIS) is known to have poor diagnosis. However, there are not enough studies that examine the biomarkers relevant to CIS development. Omics experiments generate data with tens of thousands of descriptive variables, e.g., gene expression levels. Often, many of these descriptive variables are identified as somehow relevant, resulting in hundreds or thousands of relevant variables for building models or for further data analysis. We analyze one such dataset describing patients with bladder cancer, mostly non-muscle-invasive (NMIBC), and propose a novel approach to feature selection. This approach returns high-quality features for prediction and yet allows interpretability as well as a certain level of insight into the analyzed data. As a result, we obtain a small set of seven of the most-useful biomarkers for diagnostics. They can also be used to build tests that avoid the costly and time-consuming existing methods. We summarize the current biological knowledge of the chosen biomarkers and contrast it with our findings.

Keywords: nonmuscle-invasive bladder cancer (NMIBC); carcinoma in situ (CIS); biomarker identification; optimal feature set selection

check for updates

Citation: Piliszek, R.; Brożyna, A.A.; Rudnicki, W.R. Computational Analysis Identifies Novel Biomarkers for High-Risk Bladder Cancer Patients. *Int. J. Mol. Sci.* **2022**, *23*, 7057. https://doi.org/10.3390/ijms23137057

Academic Editors: Thorsten Ecke, Thomas Otto and Florence L. Le-Calvez Kelm

Received: 26 April 2022
Accepted: 26 May 2022
Published: 24 June 2022

1. Introduction

According to the World Cancer Research Fund International, bladder cancer is the 10th most common cancer in the world [1]. It is diagnosed mostly in people over 55 in highly developed countries of southern and western Europe, as well as in North America. Men are more than four times more likely to develop bladder cancer than women. The most commonly mentioned urinary bladder cancer risks, other than being male, are smoking cigarettes, exposure to certain chemicals (such as aromatic amines, polycyclic aromatic hydrocarbons, and chlorinated hydrocarbons and alcohol), having a red meat-rich diet, and being genetically predisposed (reviewed in [1]). Urothelial carcinoma of the bladder is divided into two major groups on the basis of clinical staging with different clinical outcomes and therapy options: non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). MIBCs are aggressive tumors, characterized by a five-year survival rate of less than 50% [2]. Up to 15% of MIBCs are initially diagnosed as NMIBCs that progressed into MIBCs [3]. NMIBC is considered a tumor with a relatively good prognosis since the five-year overall survival rate is about 90% [4]. Unfortunately, NMIBCs are a very heterogeneous tumor group with a high rate of recurrence (up to 70%) and risk of progression to MIBC (up to 20%), despite significant improvement in the adjuvant therapies' efficacy (reviewed in [5–7]). Carcinoma in situ (CIS) belongs to this group and can be diagnosed as a primary or a recurrent tumor. CIS is associated with a poorer prognosis, a higher grade, as well as an elevated risk of recurrence and progression to MIBC [8]. The recurrence rate for CIS is 63–92%, and the progression to MIBC is 50–75%, even when the immunotreatment is applied [9,10]. The current treatment of CIS includes Bacille Calmette–Guérin (BCG) intravesical therapy, but up to 40% of NMIBC patients do not respond to this treatment. In these patients, one of the second-line treatments is cystectomy [11]. However, cystectomy causes side effects, especially in elderly patients. Recent studies identified some predictors of complications, with frailty index score among them [12,13]. Concomitant CIS is also related to a higher recurrence risk and mortality rate [14]. Thus, there is a need to develop accurate methods for the prediction of recurrence and progression in NMIBC, including CIS. Recently, the molecular markers predicting the progression of NMIBC have been identified [15]. However, their testing is based on the evaluation of methylation (GATA2 and TBX3) and mutation status (FGFR3); thus, its usefulness for routine use is rather limited due to the associated cost and labor of the tests [16]. Moreover, there are no specific markers for development of CIS in disease course (CIS-DC). Thus, more exact and accessible models should be developed, and new markers of CIS-DC should be identified.

The goal of the current study is to propose a small, clinically useful set of biomarkers that can be utilized for the stratification of bladder cancer patients into high- and low-risk classes, with respect to the development of CIS in disease course. The study is based on the dataset E-MTAB-4321, first described in [17] and deposited in the ArrayExpress database [18]. The dataset consists mostly of patients with Ta and T1 tumor stages. In the original analysis, the authors applied non-supervised learning to stratify patients into three groups using 119 genetic markers, showing that these three groups differ significantly in the risk of progressing to stage T2+. The original classification was extended in subsequent works by various authors [19–21].

The approach proposed in the current study is based on a robust protocol utilizing multiple supervised and non-supervised machine learning methods, including an extensive use of cross validation and resampling.

2. Materials and Methods

Dataset

The E-MTAB-4321 dataset, used in the study, contains clinical and RNA-seq data from 476 patients with early stage urothelial carcinoma, of whom 74 have developed CIS at a certain point of the disease course, whereas 402 were free of CIS during the study period. There are 43,204 genetic markers in this dataset, out of which 4800 have 0 variance, resulting in 38,404 markers actually carrying any information. A summary of the dataset characteristics is present in Table 1 and in Appendix A. For details on data collection, please refer to the original paper by Hedegaard et al. [17].

Table 1. Dataset characteristics. BCG—Bacillus Calmette–Guérin vaccine. PUNLMP—papillary urothelial neoplasm of low-malignant potential. CIS—carcinoma in situ (in the table as a stage of tumor when its sample was taken). More details on the dataset are available in the Appendices A and B.

Female	109	W/o cystectomy	444	W/o BCG treatment	388
Male	367	W/cystectomy	32	W/BCG treatment	88
CIS	3	High grade	192	W/o CIS in disease course	402
Ta	345	Low grade	277	W/CIS in disease course	74
T1	112	PUNLMP	7		
T2-4	16				

The analytical protocol is based on supervised feature selection (FS) and supervised classification. In our analysis, we focus on finding markers for predicting the appearance of CIS in disease course.

The following base feature-selection protocol is used. We first identify all informative variables and, therefore, reduce the dimensionality of the problem. Then, we further decrease the dimensionality by clustering similar variables. Finally, we use clusters' representatives to build machine learning models for the prediction of CIS-DC. Each step is described in detail in the following paragraphs.

In the first step, the variables that carry information about future development of CIS in disease course are identified. To this end, we use the multidimensional feature selection (MDFS) filter, which is based on the information entropy and is available as a library in R [22,23]. The informative variables are identified by computing information entropy conditioned on the knowledge of the descriptive variables and comparing it with the null distribution of information entropy conditioned on the non-informative variables. This metric is called information gain (IG). In this case, we use single-dimensional analysis, which computes maximum IG over multiple (30) random discretizations of continuous variables. The relevance is determined by a p-value threshold of 0.05 after applying Holm's correction [24].

Unlike minimal-optimal approaches to feature selection, all-relevant feature selection does not have a goal of producing the best set of features for model building. On the contrary—the goal is to preserve the information about all relevant variables so that they and their structure can be studied at will. However, this leads to higher complexity for model building and more uncertainty for tooling to discover such structures. To counter this, we used feature clustering to group similar features together.

Similarity is a concept rooted in clustering (and data analysis in general) and is a broad category. For our purpose, we use a correlation coefficient as our similarity metric; precisely, we use the Pearson's product–moment correlation coefficient ρ. However, for the purpose of applying clustering algorithms, we need a function that can be used as a proper metric—that monotonically describes the similar–dissimilar relation and outputs the penalty associated with dissimilarity. Thus, we apply the following transformation to obtain the function d:

$$d = 1 - \rho^2 \tag{1}$$

which satisfies the properties of a proper metric and describes dissimilarity as a penalty due to lack of correlation. The function d is called the dissimilarity function.

We choose hierarchical clustering as our clustering algorithm due to its property of revealing the internal clustering structure. As a method of hierarchical clustering, we evaluated Ward's minimum variance method as well as the complete linkage method. Of note here is that we applied clustering only to features—not objects nor both objects and features—unlike how clustering and biclustering algorithms are usually used.

We evaluate two ways to choose the representatives to build the classification models. The first is the most commonly applied procedure of working directly with the ranking of features as they are available from the feature-selection method: choosing top-n features with the lowest p-values. Secondly, we evaluate the effect of hierarchical clustering to N clusters and then, analogously, use the ranking to choose one representative from each cluster, basically the top-1 representative from each group.

To evaluate the marker set, we used the Random Forest [25] (RF) implementation available in R's randomForest package [26] as our target classifier. No tweaks to the default parameters were applied. We used the area under the ROC (receiver operating characteristic) curve, also known as AUROC or even AUC (area under curve), to describe the performance of each built classifier.

While evaluating the stability and generality of the above base protocol, we developed an extended procedure that we present here. We propose the use of cross validation as part of the feature-selection protocol. The entire above-mentioned procedure was run in a stratified 5-fold cross validation with 30 repeats, the direct results of which are presented in Figure 1. Essentially, we have obtained a new ranking from cross validation that allows us to apply the top-n procedure while using the count of repetitions as the quality metric. For an overview, see Figure 2.

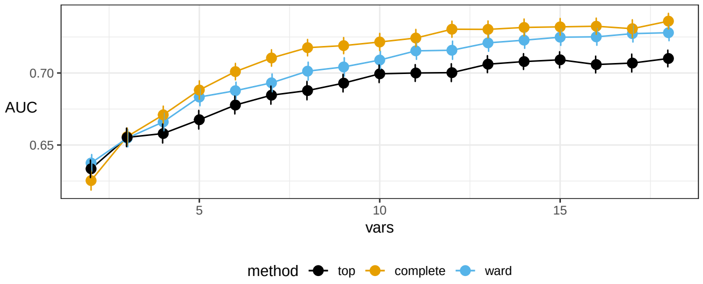

Figure 1. Plots of the area under the receiver operating characteristic curve (AUC) of the Random Forest classifiers, using markers selected by the top-*n* approach and two variants of hierarchical clustering inside our proposed protocol (complete linkage and Ward's criterion). These results were obtained without external cross validation (CV) or resampling, but from inside of the protocol itself (that is, under the protocol's internal CV). Error bars denote the standard error.

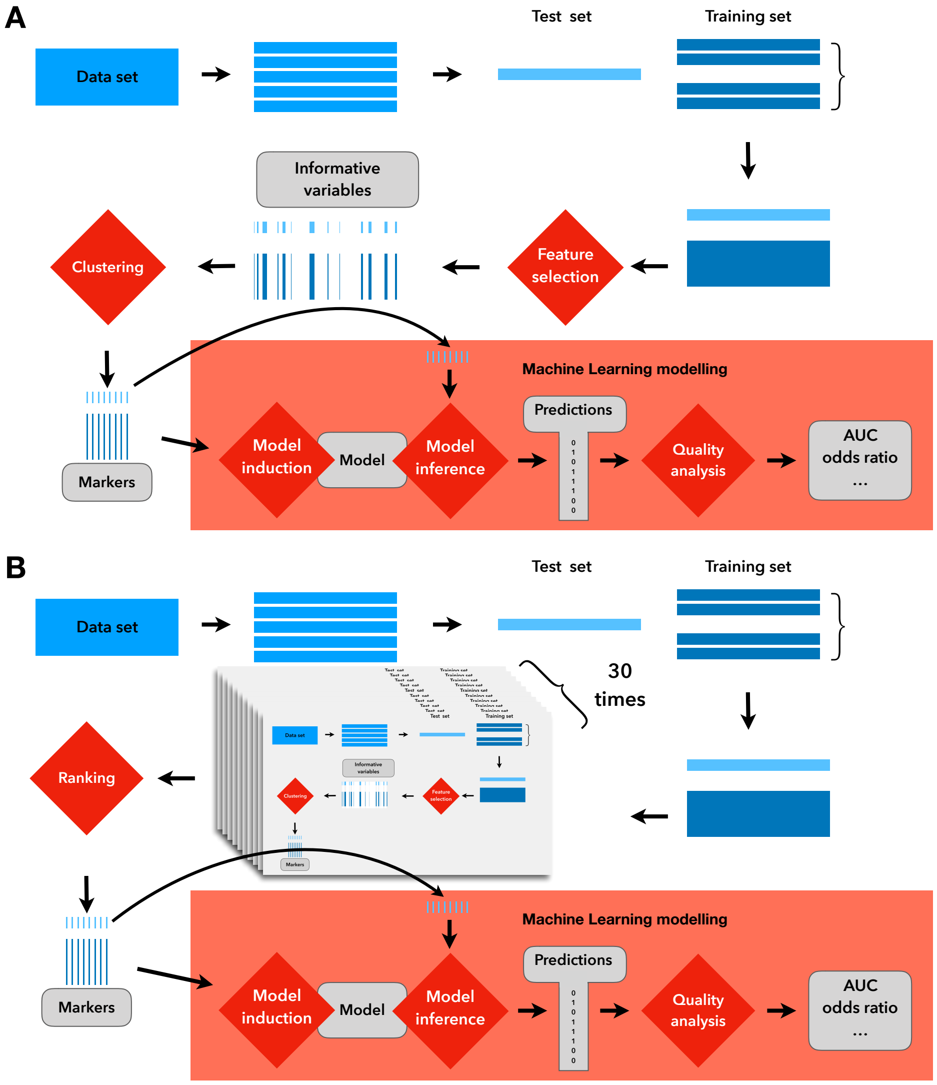

Figure 2. Depiction of a single run of the base protocol (**A**) and the proposed protocol (**B**). In (**A**), the clustering is used directly to obtain markers and build models on them. In (**B**), the (**A**) part is replicated, except for the highlighted part regarding model building and evaluation. Instead, the results of (**A**) are used to build a ranking of the most-commonly chosen variables, which then are used for model building and evaluation.

Furthermore, to estimate the mean and error of our evaluation metric (AUC), we have applied (independently) both external resampling and external cross validation. The resampling procedure consisted of 100 repeats of random sampling with replacement. The omitted objects, called out-of-bag (OOB) objects, were used for verification of the performance of the built models, i.e., for the calculation of the AUC. The cross-validation procedure, on the other hand, was conducted using a stratified 10-fold approach with 30 repeats (independent of the CV inside the procedure). The internal procedure was adjusted to use 10-fold CV as well, to gather enough objects for the MDFS statistic to work well.

Assuming validation with resampling, the full analytical protocol is, thus, as follows (with an overview in Figure 3):

1. repeat 30 times: split data randomly in 5 equal bins (i.e., run 30 repetitions of 5-fold CV) and for each (i-th) bin:
    (a) set aside the i-th bin as the test set and create a training set from the 4 remaining bins;
    (b) identify informative variables in the training set;
    (c) cluster informative variables using the hierarchical approach and select representatives of each cluster on each clustering level between 2 and 15, utilizing the usual procedure of choosing the most informative one, and:
        i. build an RF model of "CIS in disease course" using those representatives;
        ii. test the quality of the built model on the test dataset;
2. find cluster representatives at each level that appear most often in the above 150 iterations (30 times 5 iterations), at each level of clustering between 2 and 15;
3. use those representatives for building the final model on the entire dataset:
    (a) estimate the confidence intervals of the final models at each number of representatives (between 2 and 15) using the bootstrap approach—repeat 100 times:
        i. draw with replacement N patients from the original data, build RF models using the 2 to 15 representative variables;
        ii. measure the performance of each model using OOB objects;
    (b) compute the aggregate performance of each model;
    (c) use the results of the above procedure to propose the relevant markers.

Apart from the above selection of methods, we have verified the final marker set using naive Bayes [27] and logistic regression classifiers, estimating the achievable diagnostic metrics with such simpler classifiers. The details of the naive Bayes classifier are presented in the Appendices A and B, as it is used as an example simple classifier that is useful for diagnostic personnel.

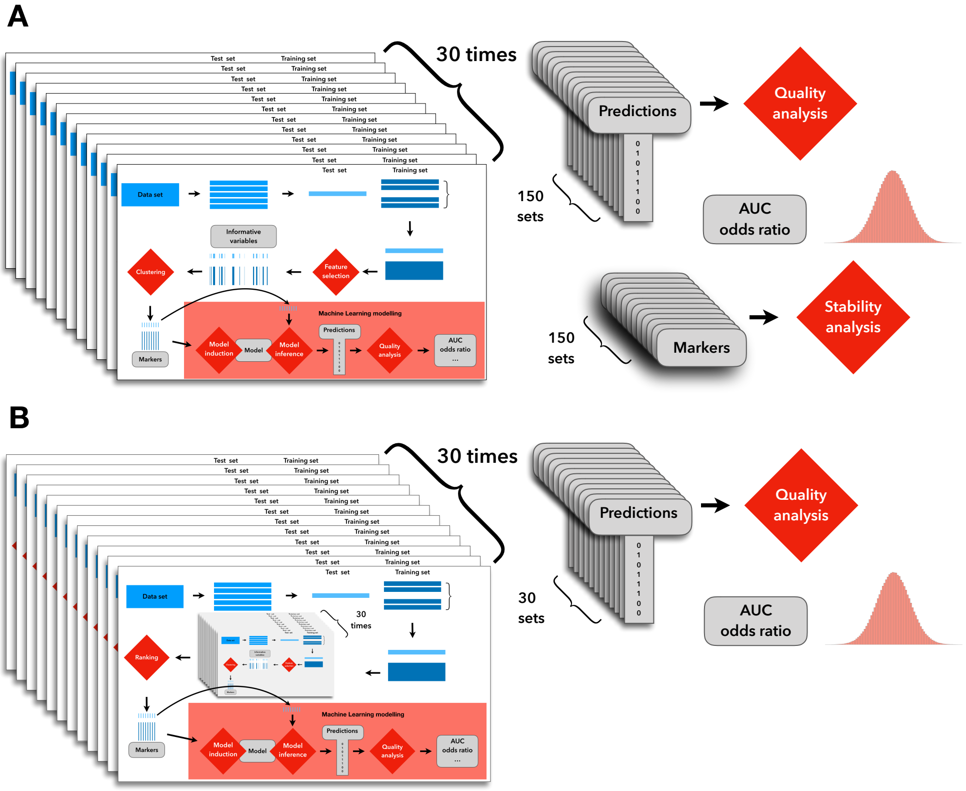

Figure 3. Depiction of the final evaluation. Both variants (**A**,**B**) from Figure 2 were evaluated, respectively. An external cross validation was used to obtain the mean and standard deviation of quality metrics (AUC, odds ratio). The evaluation of the stability of variant (**A**) (the base protocol) prompted us to create and apply variant (**B**) (the proposed protocol).

3. Results

The final set of markers was selected after inspecting both the lists of representatives and plots of the AUC in predictive models as a function of the number of markers used. The quality of the predictive models improves with the increasing number of markers used in the model, until it saturates with about seven–nine markers; see Figure 4.

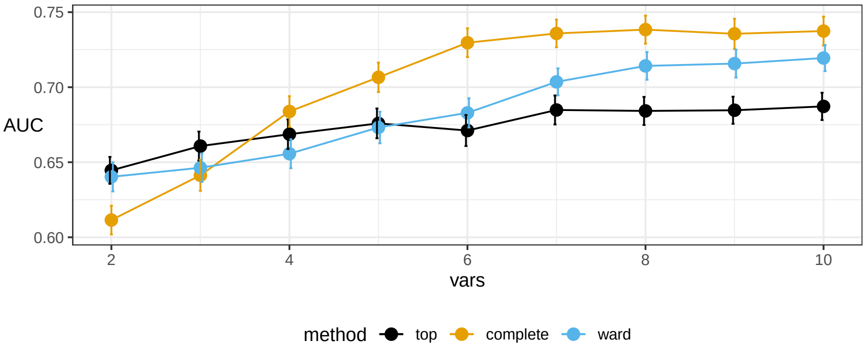

Figure 4. Plots of area under the receiver operating characteristic curve (AUC) of Random Forest classifiers using markers selected by the top-*n* approach and two variants of hierarchical clustering inside our proposed protocol (complete linkage and Ward's criterion). These results were obtained in external 10-fold cross validation (CV). Internally, for the protocol, 10-fold CV was used to ensure enough samples. Error bars denote the standard error. The complete linkage variant exhibits the desired behavior, achieving the best results earliest, with a plateau starting at 7.

An additional argument for selecting seven markers is the relative stability of the positions of the first seven markers in the list of markers consistently selected in the cross validation; see Figure 5. The first 7 markers appear in the set of the top-7 most-often selected cluster representatives in 150 repeats of the feature-selection procedure, whereas positions of other markers do not rise to the top-7.

	ENSG	Gene	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G01	169750	RAC3	G01	G02	G04	G04	G04	G04	G04	G04	G04	G04	G04	G04	G04	G04
G02	088325	TPX2	G02	G03	G03	G01	G03	G03	G03	G03	G03	G01	G01	G01	G01	G01
G03	042980	ADAM28		G01	G01	G03	G01	G01	G01	G01	G01	G03	G03	G03	G06	G06
G04	267213	DPY19L3-DT			G02	G02	G02	G05	G05	G05	G05	G06	G06	G06	G03	G03
G05	258472	E9PMD0 (paralog of SPAG5)				G05	G05	G02	G06	G06	G06	G05	G05	G05	G05	G05
G06	198720	ANKRD13B					G06	G06	G02	G02	G02	G02	G02	G02	G07	G07
G07	186952	TMEM232						G07	G07	G07	G07	G07	G07	G07	G02	G02
G08									G08	G08	G08	G08	G10	G10	G10	G08
G09										G09	G09	G10	G08	G08	G08	G10
G10											G10	G09	G09	G09	G11	G09
G11												G11	G12	G12	G09	G11
G12													G11	G11	G12	G12
G13														G13	G13	G13
G14															G14	G14
G15																G15

Figure 5. Most-representative markers at different clustering levels in 150 repeats of hierarchical clustering procedure. The first 3 columns show the order in which markers are included in the representative set, when the number of representatives is increased by 1—from 2 to 15. The Ensemble code of each marker, with 5 leading zeros removed, is shown in column 2, and the gene name corresponding to the marker is shown in column 3. In the remaining columns, the markers that are most often selected as representatives in 150 repeats are shown, and their positions within the column corresponds to the frequency of selection of a given marker as the representative (higher position—higher frequency).

The chosen markers maximize the AUC in resampling (see Figure 6) and do not exhibit strong correlations among themselves (as expected from the protocol); see Figure 7. For completeness, we also present the internal quality metric of the protocol in Figure 1 and the details of the selected seven markers in Table 2.

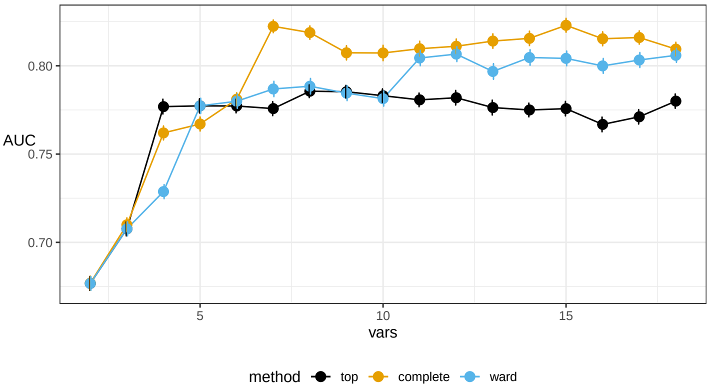

Figure 6. Plots of the area under the receiver operating characteristic curve (AUC) of Random Forest classifiers, using markers selected by the top-*n* approach and two variants of hierarchical clustering inside our proposed protocol (complete linkage and Ward's criterion). These results were obtained in 100 runs of resampling of the standard protocol, as described in the paper body. Error bars denote the standard error. The complete linkage variant again exhibits the desired behavior, achieving the best results earliest, with the plateau starting at 7.

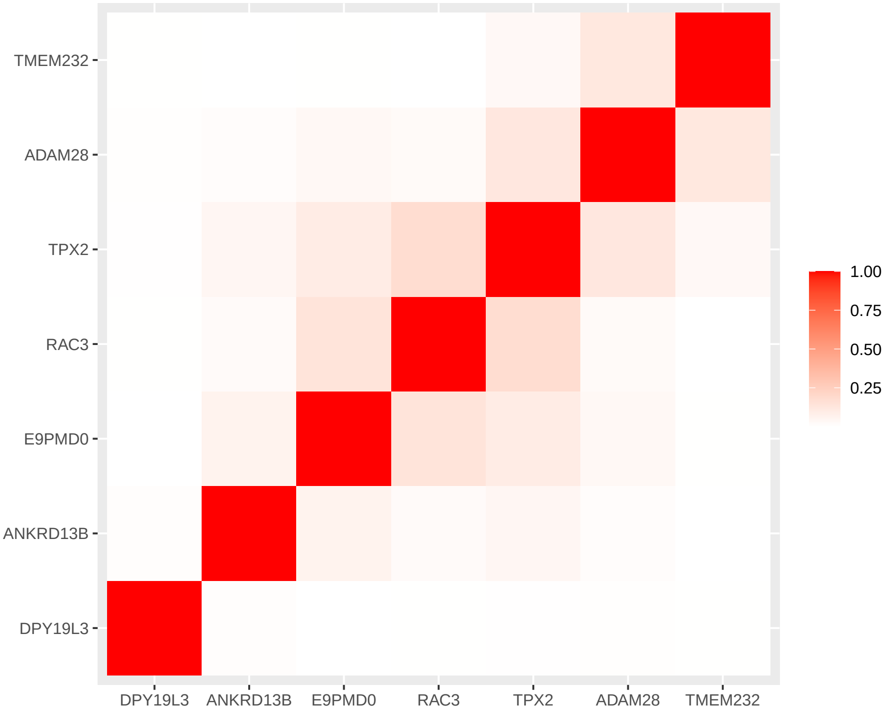

Figure 7. Heatmap of the correlation square of the chosen genes' expression levels. The darker (more saturated) the square, the higher the level of correlation.

Table 2. Details of the selected genes. The repetitions shown are for the case of selecting 7 clusters. There were 150 (30 times 5) trials, and thus, 150 is the upper bound for repetitions. IG stands for information gain, and here, it is the maximum IG computed by the MDFS library in 1D on the entire set (30 random discretizations were used). Label is a shortened version of gene name, used for identification purposes in other parts of the paper.

	Ensembl Gene ID	Repetitions	IG	Gene Name	Label
1	ENSG00000267213	114	29.3	DPY19L3-DT	DP
2	ENSG00000042980	66	35.5	ADAM28	AD
3	ENSG00000169750	64	34.0	RAC3	RA
4	ENSG00000258472	52	24.9	E9PMD0 (paralog of SPAG5)	E9
5	ENSG00000088325	50	29.0	TPX2	TP
6	ENSG00000198720	47	28.1	ANKRD13B	AN
7	ENSG00000186952	38	30.0	TMEM232	TM

The markers exhibit different directionalities of expression levels between high- and low-risk classes—ADAM28 and TMEM32 expression levels are higher in the low-risk class, while for the other markers, we observe the reverse; see Figure 8.

The diagnostic properties of the models built with the chosen markers are presented in Figure 4. Properties of models from the external cross validation are presented in Table 3. The details of the naive Bayes classifier built on the entire set are reported in Appendix A.

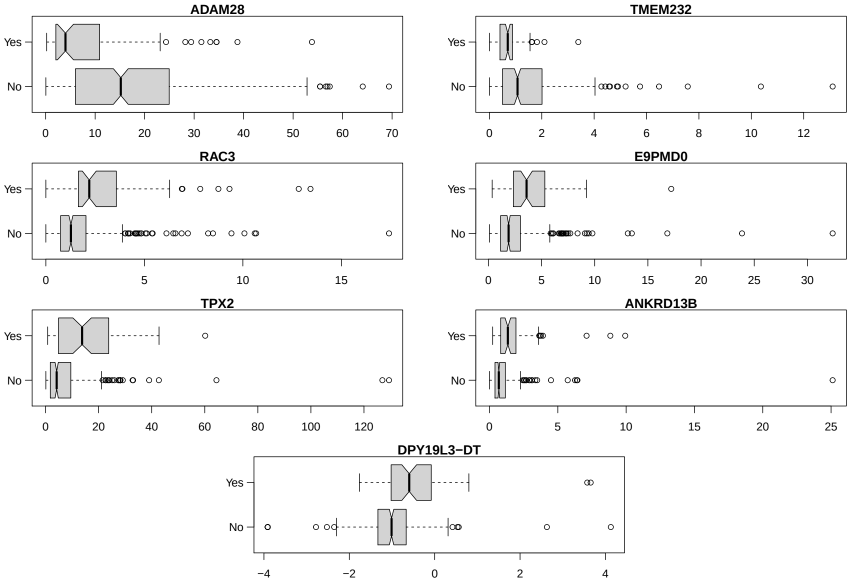

Figure 8. Boxplots of expression levels of selected markers comparing samples with CIS in disease course and those without it. Values of "DPY19L3-DT" are presented after applying a logarithm operation to be able to show the difference. Others are plotted verbatim. It can be seen that low expression levels of ADAM28 and TMEM232 increase the risk of CIS in disease course, while the 5 other variables exhibit the inverse behavior.

Table 3. Externally cross-validated results for the Random Forest classifiers. The first column defines the threshold set in the classifiers, separating low- and high-risk groups. The second column displays the fraction of all patients assigned to a high-risk class (HRC) at a given threshold. Analogously, the third column presents the fraction of all CIS-DC cases assigned to the high-risk class. Two next columns present the fraction of CIS-DC in a low- and high-risk class (LRC and HRC), respectively. Similarly, columns six and seven present the odds of CIS-DC in an LRC and HRC, respectively. Finally, the DOR column displays the diagnostic odds ratio between the HRC and LRC, and RR displays the risk ratio between these classes. The comment on the cutoff and share of patients in the HRC from Table 4 applies here as well.

Cutoff	**Share of Patients in HRC**	**Share of CIS-DC in HRC**	**Risk for LRC**	**Risk for HRC**	**Odds for LRC**	**Odds for HRC**	**DOR**	**RR**
50%	52.0%	78.8%	6.9%	23.6%	0.07	0.31	4.2	3.4
75%	23.6%	49.9%	10.2%	32.9%	0.11	0.49	4.3	3.2
90%	8.3%	24.6%	12.8%	46.2%	0.15	0.86	5.9	3.6

Table 4. Cross-validated results for the three classifiers, built using the seven selected markers. Stratified 5-fold cross validation, repeated 30 times, was used. Column headings are as in Table 3. Please note that the observed inconsistency between the cutoff and the share of patients in the HRC stems from the application of cross validation—the share is averaged over all folds from all iterations, while the cutoff is established using the entire dataset a priori.

Cutoff	Share of Patients in HRC	Share of CIS-DC in HRC	Risk for LRC	Risk for HRC	Odds for LRC	Odds for HRC	DOR	RR
Naive	Bayes							
50%	49.6%	79.5%	6.3%	24.9%	0.07	0.33	4.9	3.9
75%	25.5%	57.7%	8.8%	35.2%	0.10	0.54	5.6	4.0
90%	11.0%	33.9%	11.5%	47.9%	0.13	0.92	7.0	4.1
Logistic	Regression							
50%	49.9%	86.0%	4.4%	26.8%	0.05	0.37	8.0	6.2
75%	25.1%	57.4%	8.8%	35.5%	0.10	0.55	5.7	4.0
90%	10.6%	33.6%	11.5%	49.1%	0.13	0.97	7.4	4.3
Random	Forest							
50%	49.1%	86.6%	4.1%	27.4%	0.04	0.38	8.9	6.7
75%	26.0%	68.0%	6.7%	40.6%	0.07	0.68	9.5	6.0
90%	10.2%	35.5%	11.2%	54.3%	0.13	1.19	9.5	4.9

4. Discussion

Bladder cancer is one of the most-common cancers in the world [1]; thus, there is a need to develop sensitive methods for the early diagnosis of non-advanced lesions or poor prognosis predictors. Currently, there is a limited number of commercially available tests for bladder cancer diagnosis. The NMP22BC test allows for the diagnosis of non-muscle-invasive bladder cancer and low-grade bladder cancer in urine samples [28]. Recently published data shows that HPLC (high-performance liquid chromatography) of urine could distinguish bladder cancer patients from non-malignant hematuria patients based on chromatographic absorptions and fluorescence peaks [29]. Similarly, fluorescence urine analysis using concentration matrices of synchronous spectra could be useful in bladder cancer diagnosis, allowing to distinguish between cancer patients and heumaturia patients [30]. The new diagnostic strategy could include a label-free optical sensing platform based on DNA strand displacement. Currently, there are no data on bladder cancer detection using this method [31]. Metabolomic analysis is a very promising and useful tool for the identification of biomarkers; it allows for analyses of urine, blood, and tissue samples. The results enable distinguishing between MIBC and NMIBC patients [32]. The aforementioned techniques are aimed at the sensitive and early detection of urinary bladder cancer or at discriminations between MIBC and NMIBC. However, markers allowing for the identification of the risk of CIS development have still not been identified.

CIS of the urinary bladder represents the tumors with high risk of progression to MIBC and metastatic disease [8]. Some data indicate that primary CIS is diagnosed in about 1–3% of newly diagnosed bladder cancers, but some papers report about 20% primary CIS case diagnoses [33,34]. Secondary CIS (detected during follow-up) are diagnosed in about 20% of NMIBC cases [33,35]. Our method allowed for the identification of seven markers related to an increased risk of CIS-DC of urinary bladder cancers. Some of these markers are well-known molecules involved in cancer biology, but some of them are quite unique, with very limited information on their involvement in cancer development and their relationship with tumors.

We identified two markers that are characterized by limited information: DPY19L3-DT (DPY19L3 Divergent Transcript, ENSG00000267213) and E9PMD0 (ENSG00000258472). DPY19L3-DT belongs to the lncRNA class, but there is no information on the function of this molecule in normal and pathological cells and tissues, while the function of E9PMD0 is linked to the cell division and regulation of the attachment of spindle microtubules to kinetochore [36].

We also identified five other markers: ADAM28 (ENSG00000042980), Ras-related C3 botulinum toxin substrate 3 (Rac family small GTPase 3, RAC3, ENSG00000169750), targeting protein for Xenopus kinesin-like protein 2 (TPX2, ENSG00000088325), Ankrd13 family of ubiquitin-interacting motif (UIM)-containing proteins (Ankyrin repeat domain-containing protein 13B, ANKRD13B, ENSG00000198720), and TMEM232. Some of them were previously identified as potential cancer markers or targets for molecular anti-cancer therapies, bladder cancers among them.

ADAM28 belongs to the disintegrin and metalloprotease domain (ADAM) family. Its role in cancers is ambivalent: it promotes cancer cells' proliferation, survival, migration, and metastasis by affecting neoangiogenesis, epithelial-to-mesenchymal transition, and extracellular matrix degradation, but in the tumor microenvironment it shows strong protective effects against deleterious metastasis dissemination [37]. In bladder cancers, ADAM28 may represent a possible biomarker, since it is overexpressed in bladder transitional cell carcinoma patients and detected in urine [38,39]. In our model, its higher expression was found in patients with low-risk cancers.

Another marker identified by our protocol, RAC3, is involved in neuronal development and in tumor progression, by modulating the organization of the cytoskeleton, cell migration, cell proliferation, and reactive oxygen species production. Its expression was found in different cancers, and it is considered as a marker of poor prognosis, metastasis, and a target for molecular-targeted therapies in some human cancers, such as breast or lung (reviewed in [40]. In our model, the increased expression of RAC3 in high-risk cancers is in line with the existing knowledge and data published by Chen et al. [41]. It indicates that, in bladder cancer, this molecule can be a potential prognostic marker and a target for molecular medicine.

TPX2 is a microtubule-associated protein, involved in the assembly of mitotic spindles and in cell cycles, cell proliferation, and apoptosis [42,43]. TPX2 was found in in silico studies to be related to the risk of the distant metastasis of breast cancers [44]. In bladder cancer, TPX2 is involved in TPX2-mediated phosphorylation of the AURKA-PI3K-AKT axis [45]. In addition, heterogeneous nuclear ribonucleoprotein F, by regulating the TPX2 protein, promotes the cell cycle and proliferation of bladder cancer cells [46]. The proliferation of bladder cancer cells can also be regulated by the interplay between TPX2, p53, and GLIPR1 [47]. In our model, similar to Yan et al. [48], a higher expression of TPX2 was found in high-risk cancers. Thus, we conclude that TPX2 plays an important role in the progression of bladder cancers, including CIS in disease course, and represents a good potential marker for targeted therapy.

ANKRD13B is ubiquitin-binding protein that specifically recognizes and binds Lys-63-linked ubiquitin and that is responsible for the internalization of ligand-activated EGFR [49]. In addition, it is involved in DNA methylation since ANKRD13B (and ANKRD13A and ANKRD13D) form a complex with RNF11 (RING finger protein 11), belonging to the Really Interesting New Gene E3 ligase family (RING) [49,50]. Based on our data, we suggest that ANKRD13B could act as a marker of high-risk bladder cancer, since its expression was significantly elevated in these cancers. It could also be a potential molecular target for anticancer therapies.

TMEM232 is a member of the transmembrane protein family (TMEMs), consisting of more than 300 proteins, being components of cellular membranes [51]. Proteins of this family have differential expression in cancers, but there is limited information on TMEM232. Published data have linked this protein with atopic dermatitis [52,53] or with multiple sclerosis [54]. In our model, the TMEM232 expression pattern was similar to ADAM28, with higher expression in low-risk cancers.

Using externally cross-validated results for the Random Forest classifier and a 75% threshold in our model (Table 3), the fraction of all patients assigned to a high-risk group was 23.6%, and the fraction of all CIS-DC cases assigned to the high-risk group was 49.9%, while the fraction of CIS-DC in a low-risk group was 10.2%, and that in a high-risk group was 32.9%. The fraction of these patients for the 75% threshold, using cross validation and naive Bayes, logistic regression, and Random Forest classifiers are similar, with very promising diagnostic results for Random Forest. The described method could aid clinicians in identifying high-risk bladder cancer (the risk of CIS in disease course). Thus, it offers a diagnostic tool that allows for the personalization of bladder cancer surveillance, more precise treatment option determinations, and the improvement of bladder cancer prognoses.

To summarize, the identified genes can be used as markers of progression in urinary bladder cancers. Moreover, the increased expression of some identified proteins (RAC3, TPX2, ANKRD13B, and TMEM232) indicates their usefulness as potential targets in molecular-tailored therapies. Some of them require more detailed studies since their biological role, especially in cancer, is unknown, or the data are contradictory (ADAM28, TMEM232, DPY19L3-DT, and E9PMD0). We also conclude that, since we identified seven important genes, their evaluation in routine diagnostic procedures is possible using immunohistochemistry or in situ hybridization. Such a panel would not burden laboratories with high costs and labor. Finally, a ready classifier based on naive Bayes technique is presented in the Appendices A and B, along with an example calculation to enable the research and diagnostics communities to readily analyze applicable data.

Author Contributions: R.P. conceptualized the study, prepared the resources and the software, validated the methodology, performed the data curation and investigation, prepared the visualization of the results, and wrote, revised, and edited the manuscript; A.A.B. validated the methodology, performed formal analysis and investigation, and wrote, revised, and edited the manuscript; W.R.R. conceptualized the study, validated the methodology, performed the investigation, prepared the visualization of the results, wrote, revised, and edited the manuscript, and supervised and administrated the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In the paper, we used a publicly available dataset E-MTAB-4321 available at https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4321/.

Acknowledgments: Computations were performed at the Computational Centre at the University of Białystok, grant GO-005.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript (sorted alphabetically):

AUC	Area under the ROC curve
BCG	Bacillus Calmette–Guérin vaccine
CIS	Carcinoma in situ
CIS-DC	CIS in disease course
CV	Cross validation
DOR	Diagnostic odds ratio
FS	Feature selection
HRC	High-risk class (in this paper: predicted as with CIS-DC)
LRC	Low-risk class (in this paper: predicted as without CIS-DC)
MDFS	Multidimensional feature selection
MIBC	Muscle-invasive bladder cancer

NMIBC	Nonmuscle-invasive bladder cancer
OOB	Out-of-bag (in model training)
PUNLMP	Papillary urothelial neoplasm of low-malignant potential
RF	Random Forest
ROC	Receiver operating characteristic
RR	Risk ratio

Appendix A. Dataset Properties

In this short appendix, we summarize the details of the dataset properties in terms of the clinical metadata values' distribution. The details are in the following four tables.

Table A1. Dataset characteristics per sex (female/male). BCG—Bacillus Calmette–Guérin vaccine. PUNLMP—papillary urothelial neoplasm of low-malignant potential. CIS—carcinoma in situ (in the table as a stage of tumor when its sample was taken).

	Female	**Male**
Total	109	367
CIS-DC	18	56
Cystectomy	11	21
BCG treatment	18	70
High grade	46	146
Low grade	60	217
PUNLMP	3	4
CIS	1	2
Ta	80	265
T1	22	90
T2-4	6	10

Table A2. Dataset characteristics per CIS in disease course (CIS-DC; yes/no). BCG—Bacillus Calmette–Guérin vaccine. PUNLMP—papillary urothelial neoplasm of low-malignant potential. CIS—carcinoma in situ (in the table as a stage of tumor when its sample was taken).

	CIS-DC	**No CIS-DC**
Total	74	402
Female	18	91
Cystectomy	9	23
BCG treatment	35	53
High grade	41	151
Low grade	32	245
PUNLMP	1	6
CIS	3	0
Ta	46	299
T1	22	90
T2-4	3	13

Table A3. Dataset characteristics per cystectomy (yes/no). BCG—Bacillus Calmette–Guérin vaccine. PUNLMP—papillary urothelial neoplasm of low-malignant potential. CIS—carcinoma in situ (in the table as a stage of tumor when its sample was taken).

	Cystectomy	**No Cystectomy**
Total	32	444
Female	11	98
CIS-DC	9	65
BCG treatment	0	88
High grade	27	165
Low grade	5	272
PUNLMP	0	7
CIS	0	3
Ta	8	337
T1	18	94
T2-4	6	10

Table A4. Dataset characteristics per BCG treatment (yes/no). BCG—Bacillus Calmette–Guérin vaccine. PUNLMP—papillary urothelial neoplasm of low-malignant potential. CIS—carcinoma in situ (in the table as a stage of tumor when its sample was taken).

	BCG Treatment	**No BCG Treatment**
Total	88	388
Female	18	91
CIS-DC	35	39
Cystectomy	0	32
High grade	47	145
Low grade	41	236
PUNLMP	0	7
CIS	2	1
Ta	50	295
T1	36	76
T2-4	0	16

Appendix B. Naive Bayes Classifier Using the Seven Chosen Markers

The naive Bayes classifier built using the seven chosen markers is a simple classifier with diagnostically interesting properties (as shown in the main text). Thus, in this appendix, we present the details of classification made possible by the data we have used.

The naive Bayes classifier requires knowledge of the distribution of each variable. In most cases, it is assumed that the underlying distribution is normal. This is not the case in the raw gene expression data. Nevertheless, the logarithm of gene expression is usually sufficiently close to normal distribution. To avoid numerical artifacts for cases with very low expression, a value of 0.001 has been added to each recorded value.

To make the model general, the expression levels have been normalized using values of expression levels of genes with a stable and high level of expression that are available in the dataset. The three genes selected as reference are: ENSG00000075624 (ACTB), ENSG00000166794 (PPIB), and ENSG00000149273 (RPS3).

The values of these three have been combined to create a reference gene. The weights applied are ratios of the mean to the standard deviation, as was used in the initial choice. They are the following: 14.1, 12.8, and 8.79, respectively, normalized to 0.395, 0.357, and 0.247, respectively. The reference gene's expression level is used as the denominator in the construction of normalized expression levels of diagnostic genes, while their expression levels are the respective nominators. Thus, for a particular gene and patient, we have the following formula:

$$\frac{\log v}{0.395 \log v_{r_1} + 0.357 \log v_{r_2} + 0.247 \log v_{r_3}} \tag{A1}$$

where v is the patient's gene expression level that we normalize, and v_{r_1}, v_{r_2} and v_{r_3} are the patient's gene expression levels for reference genes mentioned above: ACTB, PPIB, and RPS3, respectively. The log function is a natural logarithm, i.e., a logarithm with base e.

Let us assume that some patient has the following values of gene expression for the respective reference genes: 490, 180, 304, and we want to normalize the value of TPX2, for which this same patient has a gene expression value of 2.03. We substitute the values in the above formula:

$$\frac{\log 2.03}{0.395 \log 490 + 0.357 \log 180 + 0.247 \log 304} \approx \frac{0.708}{2.45 + 1.85 + 1.41} = \frac{0.708}{5.71} \approx 0.124 \quad \text{(A2)}$$

The normal distribution is described using two parameters which are real numbers: the mean μ and the standard deviation (SD) σ. The density function $f_{\mu,\sigma}$ of the normal distribution with the declared parameters is expressed as follows:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{(A3)}$$

The parameters for the selected seven genes are presented per class in Table A5.

Table A5. Parameters of the normal distributions of the values of the selected markers in the used dataset, per class. SD means standard deviation. For formatting, markers are identified with the labels from Table 2.

	DP	AD	RA	E9	TP	AN	TM
Mean LRC	−0.19	0.41	0.02	0.11	0.24	−0.08	−0.02
SD LRC	0.15	0.20	0.21	0.13	0.19	0.17	0.19
Mean HRC	−0.09	0.27	0.13	0.21	0.41	0.04	−0.11
SD HRC	0.16	0.21	0.20	0.12	0.17	0.13	0.18

The name—naive Bayes—stems from the underlying statistical approach. The method uses Bayes' theorem, with an assumption that the different variables are independent. This assumption is false in most real-life cases; hence, the method is called naive. Nonetheless, the naive Bayes classifiers work surprisingly well in many cases [55].

Having the distributions well defined, equipped with the independence assumption, and using the Bayes' theorem, it is possible to evaluate whether a particular sample belongs with higher probability to the low- or high-risk class. To this end, one needs to calculate the simplified nominator from the Bayes' theorem's formula:

$$p(C_k) \prod_i p(x_i | C_k) \quad \text{(A4)}$$

where $p(C_k)$ is the prior probability of the class, while $p(x_i|C_i)$ is the conditional probability of sample with value x_i of the i-th variable, under the assumption that the sample belongs to class C_k—this is performed using the assumed distributions. The $\prod$ symbol means product. The class with the larger value is the more probable class. However, it is possible to use the classifier with an arbitrary threshold, which enables the classifier to be tweaked depending on the nature of the problem.

For practical purposes, in the case of binary classification, it is possible to define a score as a logarithm of the ratio of the probabilities of each class, in our case low-risk (L) and high-risk (H):

$$\sum_i \log \frac{p(x_i|C_L)}{p(x_i|C_H)} + \log \frac{p(C_L)}{p(C_H)} \quad \text{(A5)}$$

Since the second part is effectively a constant, and we are allowing arbitrary thresholds, the score becomes the following:

$$\text{Score} = \sum_i \log \frac{p(x_i|C_L)}{p(x_i|C_H)} \tag{A6}$$

Let us assume that the patient for whom we want to apply the method above has the following values of normalized expression levels: DP = −0.2, AD = 0.51, RA = 0.02, E9 = 0.08, TP = 0.18, AN = 0.0, and TM = 0.02. Then, utilizing Equations (A3) and (A6) (as we can substitute the probability function with probability density function since we have a ratio of them) and data from Table A5, we obtain the following contributions to the final score:

$$\begin{aligned}
\text{Score} = {} & \log \frac{f_{-0.19,0.15}(-0.2)}{f_{-0.09,0.16}(-0.2)} + \log \frac{f_{0.41,0.20}(0.51)}{f_{0.27,0.21}(0.51)} + \log \frac{f_{0.02,0.21}(0.02)}{f_{0.13,0.20}(0.02)} \\
& + \log \frac{f_{0.11,0.13}(0.08)}{f_{0.21,0.12}(0.08)} + \log \frac{f_{0.24,0.19}(0.18)}{f_{0.41,0.17}(0.18)} + \log \frac{f_{-0.08,0.17}(0.0)}{f_{0.04,0.13}(0.0)} + \log \frac{f_{-0.02,0.19}(0.02)}{f_{-0.11,0.18}(0.02)} \\
& \approx \log \frac{2.65}{1.97} + \log \frac{1.76}{0.99} + \log \frac{1.90}{1.71} + \log \frac{2.99}{1.85} + \log \frac{2.00}{0.94} + \log \frac{2.10}{2.93} + \log \frac{2.05}{1.71} \\
& \approx 0.297 + 0.575 + 0.105 + 0.480 + 0.755 + -0.333 + 0.181 = 2.06
\end{aligned} \tag{A7}$$

Taking into account the thresholds computed for the three analyzed cutoffs (Table A6), the calculated score classifies the patient as low-risk, regardless of the cutoff.

Table A6. Externally cross-validated thresholds for naive Bayes classification. Scores above the threshold mean low risk, below the threshold—high.

Cutoff	**Threshold**
50%	1.435
75%	−1.066
90%	−2.676

References

1. Saginala, K.; Barsouk, A.; Aluru, J.S.; Rawla, P.; Padala, S.A.; Barsouk, A. Epidemiology of bladder cancer. *Med. Sci.* **2020**, *8*, 15. [CrossRef]
2. Knowles, M.A.; Hurst, C.D. Molecular biology of bladder cancer: New insights into pathogenesis and clinical diversity. *Nat. Rev. Cancer* **2015**, *15*, 25–41. [CrossRef] [PubMed]
3. Chen, J.; Zhang, H.; Sun, G.; Zhang, X.; Zhao, J.; Liu, J.; Shen, P.; Shi, M.; Zeng, H. Comparison of the prognosis of primary and progressive muscle-invasive bladder cancer after radical cystectomy: A systematic review and meta-analysis. *Int. J. Surg.* **2018**, *52*, 214–220. [CrossRef] [PubMed]
4. Patel, V.G.; Oh, W.K.; Galsky, M.D. Treatment of muscle-invasive and advanced bladder cancer in 2020. *CA Cancer J. Clin.* **2020**, *70*, 404–423. [CrossRef] [PubMed]
5. Kaufman, D.S.; Shipley, W.U.; Feldman, A.S. Bladder cancer. *Lancet* **2009**, *374*, 239–249. [CrossRef]
6. Shore, N.D.; Redorta, J.P.; Robert, G.; Hutson, T.E.; Cesari, R.; Hariharan, S.; Faba, Ó.R.; Briganti, A.; Steinberg, G.D. Non-muscle-invasive bladder cancer: An overview of potential new treatment options. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2021.
7. Babjuk, M.; Burger, M.; Compérat, E.M.; Gontero, P.; Mostafid, A.H.; Palou, J.; van Rhijn, B.W.; Rouprêt, M.; Shariat, S.F.; Sylvester, R.; et al. European association of urology guidelines on non-muscle-invasive bladder cancer (TaT1 and carcinoma in situ)-2019 update. *Eur. Urol.* **2019**, *76*, 639–657. [CrossRef]
8. Tang, D.H.; Chang, S.S. Management of carcinoma in situ of the bladder: Best practice and recent developments. *Ther. Adv. Urol.* **2015**, *7*, 351–364. [CrossRef]
9. Babjuk, M.; Burger, M.; Capoun, O.; Cohen, D.; Compérat, E.M.; Escrig, J.L.D.; Gontero, P.; Liedberg, F.; Masson-Lecomte, A.; Mostafid, A.H.; et al. European association of urology guidelines on non–muscle-invasive bladder cancer (ta, T1, and carcinoma in situ). *Eur. Urol.* **2021**, *81*, 75–94. [CrossRef]
10. Griffiths, T.; Charlton, M.; Neal, D.; Powell, P. Treatment of carcinoma in situ with intravesical bacillus Calmette-Guerin without maintenance. *J. Urol.* **2002**, *167*, 2408–2412. [CrossRef]

11. Lebacle, C.; Loriot, Y.; Irani, J. BCG-unresponsive high-grade non-muscle invasive bladder cancer: What does the practicing urologist need to know? *World J. Urol.* **2021**, *39*, 4037–4046. [CrossRef]
12. De Nunzio, C.; Cicione, A.; Izquierdo, L.; Lombardo, R.; Tema, G.; Lotrecchiano, G.; Minervini, A.; Simone, G.; Cindolo, L.; D'Orta, C.; et al. Multicenter analysis of postoperative complications in octogenarians after radical cystectomy and ureterocutaneostomy: The role of the frailty index. *Clin. Genitourin. Cancer* **2019**, *17*, 402–407. [CrossRef] [PubMed]
13. Cantiello, F.; Cicione, A.; Autorino, R.; Salonia, A.; Briganti, A.; Ferro, M.; De Domenico, R.; Perdonà, S.; Damiano, R. Visceral obesity predicts adverse pathological features in urothelial bladder cancer patients undergoing radical cystectomy: A retrospective cohort study. *World J. Urol.* **2014**, *32*, 559–564. [CrossRef] [PubMed]
14. Wheat, J.C.; Weizer, A.Z.; Wolf Jr, J.S.; Lotan, Y.; Remzi, M.; Margulis, V.; Wood, C.G.; Montorsi, F.; Roscigno, M.; Kikuchi, E.; et al. Concomitant carcinoma in situ is a feature of aggressive disease in patients with organ confined urothelial carcinoma following radical nephroureterectomy. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 30, pp. 252–258.
15. Van Kessel, K.E.; van der Keur, K.A.; Dyrskjøt, L.; Algaba, F.; Welvaart, N.Y.; Beukers, W.; Segersten, U.; Keck, B.; Maurer, T.; Simic, T.; et al. Molecular markers increase precision of the european association of urology non–muscle-invasive bladder cancer progression risk groups. *Clin. Cancer Res.* **2018**, *24*, 1586–1593. [CrossRef] [PubMed]
16. Pan, C.C. The value of molecular markers in classification and prediction of progression in non-muscle-invasive bladder cancer. *Transl. Androl. Urol.* **2018**, *7*, 736. [CrossRef]
17. Hedegaard, J.; Lamy, P.; Nordentoft, I.; Algaba, F.; Høyer, S.; Ulhøi, B.P.; Vang, S.; Reinert, T.; Hermann, G.G.; Mogensen, K.; et al. Comprehensive transcriptional analysis of early-stage urothelial carcinoma. *Cancer Cell* **2016**, *30*, 27–42. [CrossRef]
18. Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; Fonseca, N.A.; Petryszak, R.; Papatheodorou, I.; et al. ArrayExpress update–from bulk to single-cell expression data. *Nucleic Acids Res.* **2019**, *47*, D711–D715. [CrossRef]
19. Aine, M.; Eriksson, P.; Liedberg, F.; Sjödahl, G.; Höglund, M. Biological determinants of bladder cancer gene expression subtypes. *Sci. Rep.* **2015**, *5*, 10957. [CrossRef]
20. Sjödahl, G.; Eriksson, P.; Liedberg, F.; Höglund, M. Molecular classification of urothelial carcinoma: Global mRNA classification versus tumour-cell phenotype classification. *J. Pathol.* **2017**, *242*, 113–125. [CrossRef]
21. Kamoun, A.; de Reyniès, A.; Allory, Y.; Sjödahl, G.; Robertson, A.G.; Seiler, R.; Hoadley, K.A.; Groeneveld, C.S.; Al-Ahmadie, H.; Choi, W.; et al. A consensus molecular classification of muscle-invasive bladder cancer. *Eur. Urol.* **2020**, *77*, 420–433. [CrossRef]
22. Mnich, K.; Rudnicki, W.R. All-relevant feature selection using multidimensional filters with exhaustive search. *Inf. Sci.* **2020**, *524*, 277 – 297. [CrossRef]
23. Piliszek, R.; Mnich, K.; Migacz, S.; Tabaszewski, P.; Sułecki, A.; Polewko-Klim, A.; Rudnicki, W. MDFS: MultiDimensional Feature Selection in R. *R J.* **2019**, *11*, 198. [CrossRef]
24. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, 65–70.
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
27. Maron, M.E. Automatic indexing: An experimental inquiry. *J. ACM (JACM)* **1961**, *8*, 404–417. [CrossRef]
28. Choi, H.S.; Lee, S.I.; Kim, D.J.; Jeong, T.Y. Usefulness of the NMP22BladderChek test for screening and follow-up of bladder cancer. *Korean J. Urol.* **2010**, *51*, 88–93. [CrossRef]
29. Džubinská, D.; Zvaník, M.; Kollárik, B.; Šikurová, L. Multiple Chromatographic Analysis of Urine in the Detection of Bladder Cancer. *Diagnostics* **2021**, *11*, 1793. [CrossRef]
30. Kollarik, B.; Zvarik, M.; Bujdak, P.; Weibl, P.; Rybar, L.; Sikurova, L.; Hunakova, L. Urinary fluorescence analysis in diagnosis of bladder cancer. *Neoplasma* **2018**, *65*, 234–241. [CrossRef]
31. Zhang, Y.; Wang, L.; Wang, Y.; Dong, Y. Label-free optical biosensor for target detection based on simulation-assisted catalyzed hairpin assembly. *Comput. Biol. Chem.* **2019**, *78*, 448–454. [CrossRef]
32. Di Meo, N.A.; Loizzo, D.; Pandolfo, S.D.; Autorino, R.; Ferro, M.; Porta, C.; Stella, A.; Bizzoca, C.; Vincenti, L.; Crocetto, F.; et al. Metabolomic Approaches for Detection and Identification of Biomarkers and Altered Pathways in Bladder Cancer. *Int. J. Mol. Sci.* **2022**, *23*, 4173. [CrossRef]
33. Piszczek, R.; Krajewski, W.; Małkiewicz, B.; Krajewski, P.; Tukiendorf, A.; Zdrojowy, R.; Kołodziej, A. Clinical outcomes and survival differences between primary, secondary and concomitants carcinoma in situ of urinary bladder treated with BCG immunotherapy. *Transl. Androl. Urol.* **2020**, *9*, 1338. [CrossRef] [PubMed]
34. Nese, N.; Gupta, R.; Bui, M.H.; Amin, M.B. Carcinoma in situ of the urinary bladder: Review of clinicopathologic characteristics with an emphasis on aspects related to molecular diagnostic techniques and prognosis. *J. Natl. Compr. Cancer Netw.* **2009**, *7*, 48–57. [CrossRef] [PubMed]
35. Hayakawa, N.; Kikuchi, E.; Mikami, S.; Matsumoto, K.; Miyajima, A.; Oya, M. The clinical impact of the classification of carcinoma in situ on tumor recurrence and their clinical course in patients with bladder tumor. *Jpn. J. Clin. Oncol.* **2011**, *41*, 424–429. [CrossRef] [PubMed]
36. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef] [PubMed]
37. Hubeau, C.; Rocks, N.; Cataldo, D. ADAM28: Another ambivalent protease in cancer. *Cancer Lett.* **2020**, *494*, 18–26. [CrossRef]

38. Yang, M.H.; Chu, P.Y.; Chen, S.C.J.; Chung, T.W.; Chen, W.C.; Tan, L.B.; Kan, W.C.; Wang, H.Y.; Su, S.B.; Tyan, Y.C. Characterization of ADAM28 as a biomarker of bladder transitional cell carcinomas by urinary proteome analysis. *Biochem. Biophys. Res. Commun.* **2011**, *411*, 714–720. [CrossRef]
39. Tyan, Y.C.; Yang, M.H.; Chen, S.C.J.; Jong, S.B.; Chen, W.C.; Yang, Y.H.; Chung, T.W.; Liao, P.C. Urinary protein profiling by liquid chromatography/tandem mass spectrometry: ADAM28 is overexpressed in bladder transitional cell carcinoma. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 2851–2862. [CrossRef]
40. De Curtis, I. The Rac3 GTPase in neuronal development, neurodevelopmental disorders, and cancer. *Cells* **2019**, *8*, 1063. [CrossRef]
41. Chen, M.; Nie, Z.; Cao, H.; Gao, Y.; Wen, X.; Zhang, C.; Zhang, S. Rac3 Expression and its Clinicopathological Significance in Patients With Bladder Cancer. *Pathol. Oncol. Res.* **2021**, *27*, 28. [CrossRef]
42. Moss, D.K.; Wilde, A.; Lane, J.D. Dynamic release of nuclear RanGTP triggers TPX2-dependent microtubule assembly during the apoptotic execution phase. *J. Cell Sci.* **2009**, *122*, 644–655. [CrossRef]
43. Bird, A.W.; Hyman, A.A. Building a spindle of the correct length in human cells requires the interaction between TPX2 and Aurora A. *J. Cell Biol.* **2008**, *182*, 289–300. [CrossRef] [PubMed]
44. Cai, Y.; Mei, J.; Xiao, Z.; Xu, B.; Jiang, X.; Zhang, Y.; Zhu, Y. Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. *Hereditas* **2019**, *156*, 1–12. [CrossRef] [PubMed]
45. Li, X.; Wei, Z.; Yu, H.; Xu, Y.; He, W.; Zhou, X.; Gou, X. Secretory autophagy-induced bladder tumour-derived extracellular vesicle secretion promotes angiogenesis by activating the TPX2-mediated phosphorylation of the AURKA-PI3K-AKT axis. *Cancer Lett.* **2021**, *523*, 10–28. [CrossRef]
46. Li, F.; Su, M.; Zhao, H.; Xie, W.; Cao, S.; Xu, Y.; Chen, W.; Wang, L.; Hou, L.; Tan, W. HnRNP-F promotes cell proliferation by regulating TPX2 in bladder cancer. *Am. J. Transl. Res.* **2019**, *11*, 7035.
47. Yan, L.; Li, Q.; Yang, J.; Qiao, B. TPX2-p53-GLIPR1 regulatory circuitry in cell proliferation, invasion, and tumor growth of bladder cancer. *J. Cell. Biochem.* **2018**, *119*, 1791–1803. [CrossRef] [PubMed]
48. Yan, L.; Li, S.; Xu, C.; Zhao, X.; Hao, B.; Li, H.; Qiao, B. Target protein for Xklp2 (TPX2), a microtubule-related protein, contributes to malignant phenotype in bladder carcinoma. *Tumor Biol.* **2013**, *34*, 4089–4100. [CrossRef]
49. Mattioni, A.; Boldt, K.; Auciello, G.; Komada, M.; Rappoport, J.Z.; Ueffing, M.; Castagnoli, L.; Cesareni, G.; Santonico, E. Ring Finger Protein 11 acts on ligand-activated EGFR via the direct interaction with the UIM region of ANKRD13 protein family. *FEBS J.* **2020**, *287*, 3526–3550. [CrossRef]
50. Cho, N.Y.; Park, J.W.; Wen, X.; Shin, Y.J.; Kang, J.K.; Song, S.H.; Kim, H.P.; Kim, T.Y.; Bae, J.M.; Kang, G.H. Blood-based detection of colorectal cancer using cancer-specific DNA methylation markers. *Diagnostics* **2021**, *11*, 51. [CrossRef]
51. Wrzesiński, T.; Szelag, M.; Cieślikowski, W.A.; Ida, A.; Giles, R.; Zodro, E.; Szumska, J.; Poźniak, J.; Kwias, Z.; Bluyssen, H.A.; et al. Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. *BMC Cancer* **2015**, *15*, 518. [CrossRef]
52. Wu, Y.Y.; Tang, J.P.; Liu, Q.; Zheng, X.D.; Fang, L.; Yin, X.Y.; Jiang, X.Y.; Zhou, F.S.; Zhu, F.; Liang, B.; et al. Scanning indels in the 5q22. 1 region and identification of the TMEM232 susceptibility gene that is associated with atopic dermatitis in the Chinese Han population. *Gene* **2017**, *617*, 17–23. [CrossRef]
53. Zheng, J.; Wu, Y.Y.; Fang, W.L.; Cai, X.Y.; Yu, C.X.; Zheng, X.D.; Xiao, F.L. Confirming the TMEM232 gene associated with atopic dermatitis through targeted capture sequencing. *Sci. Rep.* **2021**, *11*, 21830. [CrossRef] [PubMed]
54. Souren, N.Y.; Gerdes, L.A.; Lutsik, P.; Gasparoni, G.; Beltrán, E.; Salhab, A.; Kümpfel, T.; Weichenhan, D.; Plass, C.; Hohlfeld, R.; et al. DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nat. Commun.* **2019**, *10*, 2094. [CrossRef] [PubMed]
55. Stephens, C.R.; Huerta, H.F.; Linares, A.R. When is the Naive Bayes approximation not so naive? *Mach. Learn.* **2018**, *107*, 397–441. [CrossRef]