




## Article

# Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data

I-Jung Tsai <sup>1,†</sup> , Wen-Chi Shen <sup>2,†</sup>, Chia-Ling Lee <sup>3</sup>, Horng-Dar Wang <sup>2,4,5,\*</sup>  and Ching-Yu Lin <sup>1,6,\*</sup> 

- <sup>1</sup> Ph.D. Program in Medical Biotechnology, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan; d609108005@tmu.edu.tw
- <sup>2</sup> Institute of Biotechnology, National Tsing Hua University, Hsinchu 30013, Taiwan; tt22407@gmail.com
- <sup>3</sup> Pathology Department, MacKay Memorial Hospital, Taipei 10449, Taiwan; g660108001@tmu.edu.tw
- <sup>4</sup> Department of Life Science, National Tsing Hua University, Hsinchu 30013, Taiwan
- <sup>5</sup> Institute of Systems Neuroscience, National Tsing Hua University, Hsinchu 30013, Taiwan
- <sup>6</sup> School of Medical Laboratory Science and Biotechnology, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan
- \* Correspondence: hdwang@life.nthu.edu.tw (H.-D.W.); cylin@tmu.edu.tw (C.-Y.L.); Tel.: +886-3-5742470 (H.-D.W.); +886-2-27361661 (ext. 3326) (C.-Y.L.)
- † These authors contributed equally to this work.

**Abstract:** Bladder cancer has been increasing globally. Urinary cytology is considered a major screening method for bladder cancer, but it has poor sensitivity. This study aimed to utilize clinical laboratory data and machine learning methods to build predictive models of bladder cancer. A total of 1336 patients with cystitis, bladder cancer, kidney cancer, uterus cancer, and prostate cancer were enrolled in this study. Two-step feature selection combined with WEKA and forward selection was performed. Furthermore, five machine learning models, including decision tree, random forest, support vector machine, extreme gradient boosting (XGBoost), and light gradient boosting machine (GBM) were applied. Features, including calcium, alkaline phosphatase (ALP), albumin, urine ketone, urine occult blood, creatinine, alanine aminotransferase (ALT), and diabetes were selected. The lightGBM model obtained an accuracy of 84.8% to 86.9%, a sensitivity 84% to 87.8%, a specificity of 82.9% to 86.7%, and an area under the curve (AUC) of 0.88 to 0.92 in discriminating bladder cancer from cystitis and other cancers. Our study provides a demonstration of utilizing clinical laboratory data to predict bladder cancer.

**Keywords:** machine learning; bladder cancer; feature selection; clinical laboratory data



**Citation:** Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y. Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data. *Diagnostics* **2022**, *12*, 203. <https://doi.org/10.3390/diagnostics12010203>

Academic Editor: Christoph Palm

Received: 23 December 2021

Accepted: 13 January 2022

Published: 14 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Bladder cancer has been noted as the 10th most common cancer in the world [1]. The incidence of bladder cancer is rising globally, especially in developed countries, such as U.S.A, Germany, and Taiwan; according to GLOBOCAN, 573,278 new cases of bladder cancer and 212,536 new deaths [2]. Furthermore, bladder cancer is observed in men more than in women, with respective incidence and mortality rates of 9.5 and 3.3 per 100,000 among men, which are four times those among women globally [2]. Moreover, smoking is considered the major risk factor in patients with bladder cancer [3]. The gold standard procedure for diagnosing bladder cancer is cystoscopy, with a sensitivity 88–100% and specificity 77.1–97% [4]. Currently, urinary cytology is considered a major non-invasive method to diagnose bladder cancer with high specificity, but only 38% sensitivity [5]. Therefore, a screening method with high sensitivity and high specificity is urgently needed for the diagnosis of bladder cancer.

Clinical chemistry tests and urinalysis are the major diagnostic screening test in the clinical laboratory [6]. The alteration of each test can be interpreted as a relationship with diseases; for instance, aspartate aminotransferase (AST) and alanine aminotransferase

(ALT) are the enzymes from liver, pancreas, and kidneys [7]. Furthermore, an increased level of AST and ALT is used in the diagnosis of liver disease; however, it is also related to the damage of other organs [7]. Large amounts of alkaline phosphatase (ALP) can be found in liver, bone, intestine, and placenta; while, ALP can be considered an indicator of bone formation [8]. In addition, the analysis of isoenzymes of ALP is an aid in diagnosing bone and liver disease [9]. Studies indicated that ALP may be related to diseases such as chronic kidney disease [10], rheumatoid disease [11], malignant disease [12], and prostate cancer [13]. The physiological meaning of the alteration in ions, such as potassium ions, sodium ions, calcium ions, and chloride, is complicated. Studies indicated that the alteration of sodium ions and calcium ions may be related to hypertension [14], alkalosis [15], cancers [16], and multiple sclerosis [17]. The serum level of albumin can be considered the sum of synthesis, degradation, and distribution [18]. In addition, albumin in the blood exhibits many biological functions, including transport of endogenous and exogenous compounds, modulation of capillary permeability, neutrophil homeostasis, and free radical scavenging [19]. The alteration of serum albumin level has been found to be correlated with coronary heart disease [20], bowel disease [21], and liver disease [19]. Whereas the urine albumin has been reported to be associated with chronic kidney disease [22] and heart disease [23]. Creatinine is the amino acid compound from creatine and is released into the blood mainly from muscles; meanwhile, the creatinine in the blood is released constantly and filtered by glomerulus in the kidney [24]. Furthermore, the creatinine in blood can be used to calculate the glomerular filtration rate, to evaluate the function of the kidney [25]. Therefore, the alteration of creatinine in the blood may be related to muscle disease and kidney disease [26]. A complete urinalysis consists of color, clarity, specific gravity, and chemical analyses, as well as examination of urine sediment [27]. A chemical analysis is usually conducted to examine pH, glucose, ketones, occult blood, bilirubin, and protein with dipstrip systems [28]. Aberrant results in chemical analysis can occur in certain diseases; for instance, ketonuria can be found in the urine from patients with diabetes; while, hematuria indicates bleeding within the urinary tract [29]. Moreover, the examination of urine sediment may be found with crystal, erythrocytes, leukocytes, bacteria, and casts, which provide information about the urinary tract system [30]. However, interpreting clinical tests individually may cause a misleading diagnosis [31]. The laboratory test results can be interpreted with experienced clinicians, but it can also be integrated and interpreted with artificial intelligence (AI), such as machine learning algorithms [32].

Machine learning is a type of AI whereby computers independently learn from data, without human intervention [33]. Furthermore, machine learning has different methods to approach the learning process [34]. Various machine learning algorithms have been frequently used in medical studies. Tree based models, such as decision tree, random forest, extreme gradient boosting (XGBoost), and light gradient boosting machine (GBM) have been frequently used in medical studies [35–37]. Moreover, support vector machine has been considered an alternative approach for managing clinical data [38]. Recently, many studies in machine learning applications have been introduced into clinical practice. Machine learning has become a powerful tool for improving diagnostic and prognostic accuracy in cancer, with various kinds of data; for instance, Cai et al. exploited genomic data to create a panel of 16 DNA methylation markers and combined them with random forest to provide a classification power with an accuracy of 86.54% in classifying lung cancer [39]. Routine clinical and laboratory data were used to establish machine learning models to identify lung cancer in an early stage [40]. An elegant study from Obaid et al. demonstrated the potential of integrating image data with machine learning to diagnose breast cancer and received an accuracy of 98.1% [41]. As for bladder cancer, Garapati et al. built an objective computer-aided system to identify the stage of bladder cancer with CT urography [42]. Additionally, machine learning was also applied to metabolomics to recognize early and late stages of bladder cancer [43]. However, we found that no study has applied machine learning with clinical laboratory data for improving the diagnostic accuracy of patients with bladder cancer.

Herein, we collected clinical laboratory data, including biochemistry tests and urinalysis from 1336 patients with cystitis, bladder cancer, and other types of cancer in Mackay Memorial Hospital. We combined sampling techniques and two-step feature selection to exploit the clinical laboratory dataset. Furthermore, five different machine learning models were trained and validated with the selected dataset. Moreover, the accuracy, precision, f1 score, sensitivity, specificity, and area under the area of receiving operating characteristic curve (AUC) were calculated to evaluate the model performance.

## 2. Materials and Methods

### 2.1. Patient Cohort

We collected clinical laboratory test data from 144 patients with cystitis (56 female and 88 male patients, aged  $60.12 \pm 11.99$ ), 200 patients with kidney cancer (62 female and 138 male patients, aged  $63.41 \pm 10.45$ ), 201 patients with prostate cancer (201 male patients, aged  $71.83 \pm 6.42$ ), 591 patients with bladder cancer (205 female and 386 male patients, aged  $66.73 \pm 9.4$ ), and 200 patients with uterus cancer (200 female patients, aged  $60.86 \pm 10.26$ ) in Mackay Memorial Hospital from January 2017 to February 2020. Patients with cancers were diagnosed by clinician and confirmed via pathological report. MacKay Memorial Hospital Institutional Review Board approved the study protocol (20MMHIS200e (8 July 2020)).

### 2.2. Statistical Analysis

Description analyses were performed with SPSS 19.0 (IBM, Chicago, IL, USA). Continuous variables are presented as mean  $\pm$  SD or median (25th and 75th percentile). The categorical variables are presented as number and percentage. The clinical laboratory test and characteristics were compared with a *t*-test or Mann–Whitney U test for continuous variables and chi-square test for categorical variables.

### 2.3. Data Processing

We collected clinical laboratory data with 56 laboratory tests results. The laboratory test results with more than 50% of missing data were removed. After that, we received 31 laboratory test results with missing rate, varying between 0 to 44.1% (Table 1). The missing data was filled with the mean value for continuous value and median value for categorical value from each feature in the whole data. Features that were missing in the data in certain classifications were avoided in the feature selecting, model training, and validating. For instance, A/G ratio and urine epithelium was not included while discriminating cystitis from other cancers. The oversampling and undersampling techniques from imblearn v0.0 package were used for the problem of imbalanced data [44].

### 2.4. Feature Selection and Machine Learning

We used an InfoGainAttributeEval (InfoGain) + Ranker method with default parameters to perform feature selection with WEKA (vers. 3.8.3) (Table 2). Furthermore, optimized models were used to conduct a forward selection, as mentioned in a previous study [45]. The models we built in this study were based on decision tree (DT), random forest (RF), support vector machine (SVM), XGBoost, and lightGBM, with 10-fold cross validation with scikit-learn (vers. 0.21.3). The parameters were tuned before the experiments. For DT, the initial value of tree depth was set from 1–10, with a step of 1. The kernel of the model was set to entropy or gini. For RF, the initial value of the tree number was set at 100 and increased by 100 until 500. The kernel of model was set to gini or entropy. For SVM, the initial value of gamma was set from  $10^{-6}$  to  $10^{-10}$ , with a step of 0.1. The initial value of C was set from  $10^{-6}$  to  $10^{-7}$  with a step of 10. The kernel of SVM was set to RBF. For XGBoost, the initial value of eta was from 0.01 to 0.2, with a step of 0.05. The initial value of depth was set from 1 to 10, with a step of 1. As for lightGBM, the initial value of leaves was set from 50 to 400, with a step of 50. The initial value of depth was set from 1 to 10, with a step of 1. The parameters of the machine learning were tuned via training and validating with the whole dataset. The parameters that obtained the highest accuracy

were selected (Table 2). A confusion matrix was used in this study to calculate the accuracy, precision, sensitivity, specificity, and f1 score (Table 3). The value of AUC was calculated from scikit-learn (vers. 0.21.3).

**Table 1.** The missing data rate in the clinical laboratory data.

Feature	Data Type	Missing Data	Missing Data (%)
A/G Ratio	Continuous	0.441	44.1
Albumin	Continuous	0.224	22.4
ALP	Continuous	0.378	37.8
ALT	Continuous	0.068	6.8
AST	Continuous	0.046	4.6
BUN	Continuous	0.018	1.8
Calcium	Continuous	0.428	42.8
Chloride	Continuous	0.084	8.4
Creatinine	Continuous	0.005	0.5
Direct Bilirubin	Continuous	0.303	30.3
Estimated GFR	Continuous	0.008	0.8
Glucose AC	Continuous	0.038	3.8
Nitrite	Categorical	0.026	2.6
Urine occult Blood	Categorical	0.026	2.6
pH	Continuous	0.026	2.6
Potassium	Continuous	0.035	3.5
Sodium	Continuous	0.037	3.7
Specific Gravity	Continuous	0.026	2.6
Strip WBC	Continuous	0.16	16
Total Bilirubin	Continuous	0.216	21.6
Total Cholesterol	Continuous	0.19	19
Total Protein	Continuous	0.285	28.5
Triglyceride	Continuous	0.204	20.4
Urine epithelium (UL)	Continuous	0.43	43
Urine epithelium count	Continuous	0.02	2
Uric acid	Continuous	0.15	15
Urine Bilirubin	Categorical	0.026	2.6
Urine Glucose	Categorical	0.16	16
Urine Ketone	Categorical	0.16	16
Urine Protein	Categorical	0.026	2.6
Urobilinogen	Categorical	0.026	2.6

A/G Ratio: albumin globulin ratio. BUN: blood urea nitrogen.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (4)$$

$$\text{f1 score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{sensitivity}}} \quad (5)$$



**Table 2.** The parameters of machine learning and feature selection.

Algorithm Name	Parameter Name	Parameter Value
InfoGainAttributeEval	binarizeNumericAttributes	False
	doNotCheckCapabilities	False
	missingMerge	True
Ranker	generateRanking numToSelect	True −1
Decision tree	criterion of tree depth of tree	gini 4
Random forest	criterion of tree estimators	gini 300
SVM	kernel C value gamma	rbf 1000 0.000001
XGBoost	eta depth of tree	0.2 7
LightGBM	number of leaf depth of tree	100 1

**Table 3.** The confusion matrix for evaluation of model performance.

Prediction	Patients		
		Bladder Cancer	Cystitis
	bladder cancer	true positive, TP	false positive, FP
cystitis	false negative, FN	true negative, TN	

### 3. Results

#### 3.1. Clinical Characterisitcs and Clinclal Laboratory Data from Patients

The differences in demographics, baseline characteristics, and laboratory data between healthy groups and other cancers are summarized in Table 4. We found that albumin, ALP, BUN, chloride, creatinine, direct bilirubin, eGFR, pH, potassium, total protein, nitrite, strip WBC, and urine occult blood were significantly different in patients with kidney cancer compared to patients with cystitis. Furthermore, we discovered that ALP, AST, BUN, calcium, creatinine, sodium, urine epithelium counts, and urine occult blood had significant differences in patients with prostate cancer compared to patients with cystitis. As for bladder cancer, the statistical results shown that ALP, BUN, calcium, chloride, creatinine, direct bilirubin, eGFR, glucose, specific gravity, total protein, and uric acid were significantly different compared to patients with cystitis. Lastly, ALP, BUN, calcium, chloride, creatinine, eGFR, glucose, potassium, sodium, urine epithelium count, urine protein, urobilinogen, and urine occult blood were significantly different between patients with uterus cancer and patients with cystitis.

**Table 4.** Comparison of clinical characteristics and clinical laboratory data between patients with cystitis and patients with other cancers.

	Cystitis <i>n</i> = 144	Kidney Cancer <i>n</i> = 200	Prostate Cancer <i>n</i> = 201	Bladder Cancer <i>n</i> = 591	Uterus Cancer <i>n</i> = 200
age	60.12 ± 11.99	63.41 ± 10.45 **	71.83 ± 6.42 **	66.73 ± 9.4 **	60.86 ± 10.26
sex	88 (61.1%)	138 (69%)	201 (100%) **	386 (65.3%)	0 **
hypertension	34 (23.7%)	72 (36%) *	66 (32.8%)	173 (29.3%)	22 (11%) *
diabetes	20 (13.9%)	46 (23%) *	33 (16.4%)	93 (15.7%)	8 (4%) **
smoking	18 (12.5%)	51 (25.5%) *	47 (23.4%) *	138 (23.4%) *	9 (4.5%) *
drinking	23 (16%)	41 (20.5%)	64 (31.8%) **	118 (20%)	17 (8.5%) **
beetle nuts	2 (1.4%)	3 (1.5%)	3 (1.5%)	20 (3.4%)	1 (0.5%)
family history	1 (0.7%)	7 (3.5%)	5 (2.5%)	12 (2%)	7 (3.5%)

Table 4. Cont.

	Cystitis <i>n</i> = 144	Kidney Cancer <i>n</i> = 200	Prostate Cancer <i>n</i> = 201	Bladder Cancer <i>n</i> = 591	Uterus Cancer <i>n</i> = 200
A/G Ratio	-	1.75 ± 0.4	1.07 ± 0.46	1.61 ± 0.47	1.64 ± 0.38
Albumin	3.96 ± 0.64	4.27 ± 0.57 **	4 ± 0.68	4 ± 0.68	4.15 ± 0.58 *
ALP	71 (55, 91)	66 (52, 79) **	66 (60, 72) **	69 (55, 90) **	65.5 (53, 79)
ALT	30.12 ± 42.66	31.46 ± 35.99	29.94 ± 32.49	27.45 ± 31.9	25.65 ± 28.02
AST	28.31 ± 32.07	30.56 ± 29.51	43.13 ± 74.83 *	40.65 ± 293.66	29.33 ± 53.46
BUN	14 (11, 21)	16 (12, 21) **	17 (13, 22.9) *	16 (12, 26) **	12 (9, 16.85) **
Calcium	9 (8.5, 9.4)	9.3 (8.85, 9.6)	8.895 (8.3, 9.4) **	9 (8.5, 9.4) **	9.2 (8.65, 9.65) **
Chloride	105.2 (103, 107.5)	105 (103, 107) *	105 (102.075, 107.225)	105 (102, 108) *	106 (104, 108) **
Creatinine	0.9 (0.7, 1.2)	1.1 (0.9, 1.5) **	1 (0.9, 1.2) *	1.1 (0.8, 1.5) **	0.7 (0.6, 0.8) **
Direct Bilirubin	0.11 (0.1, 0.2)	0.1 (0.1, 0.2) **	0.2 (0.1, 0.2)	0.1 (0.1, 0.2) *	0.1 (0.1, 0.18)
Estimated GFR	75.38 ± 33.22	62.68 ± 32.54 **	71.92 ± 27.16	64.13 ± 34 **	91.57 ± 28.86 **
Glucose AC	120.65 ± 47.39	123.11 ± 41.14	129.36 ± 56.6	124.83 ± 93.83	109.57 ± 26.12 *
pH	6.24 ± 0.89	6.03 ± 0.74 *	6.09 ± 0.85	6.26 ± 0.86	6.12 ± 0.71
Potassium	3.92 ± 0.5	4.14 ± 0.52 **	4.02 ± 0.48	4.12 ± 0.61 **	4.06 ± 0.46 **
Sodium	139 (137, 141)	139 (137, 140.775)	138 (136, 140) *	139 (137, 140.8)	139 (138.5, 141) **
Specific Gravity	1.016 (1.012, 1.021)	1.017 (1.013, 1.021)	1.017 (1.011, 1.022)	1.014 (1.01, 1.0195) *	1.016 (1.011, 1.021)
Total Bilirubin	0.9 ± 0.49	0.84 ± 0.31	1.08 ± 1.25	1.04 ± 1.93	0.81 ± 0.68
Total Cholesterol	190.15 ± 49.57	182.2 ± 42.36	189.88 ± 45.34	183.76 ± 44.58	200.1 ± 50.31
Total Protein	6.6 (6.0625, 7)	6.9 (6.55, 7.2) **	6.7 (6, 7.1)	6.7 (6.1, 7.1325) **	6.9 (6.3, 7.3)
Triglyceride	143.47 ± 88.07	141.88 ± 170.99	142.3 ± 89.64	131.97 ± 91.15	133.65 ± 141.94
Uric acid	5.82 ± 1.83	6.1 ± 1.64	6.36 ± 2.61	6.22 ± 1.88 *	5.4 ± 1.76
Urine epithelium (UL)	-	0 (0, 6)	2.5 (0, 5.75)	3 (0, 9)	-
Urine epithelium count	0 (0, 2)	0 (0, 2)	0 (0, 2) **	1 (0, 3)	2 (0, 3) *
Nitrite		<i>p</i> = 0.001			
0	127 (88.2%)	194 (97%)	188 (93.5%)	493 (83.4%)	181 (90.5%)
1	17 (11.8%)	6 (3%)	13 (6.5%)	98 (16.6%)	19 (9.5%)
Strip WBC		<i>p</i> = 0.001	<i>p</i> = 0.003		
0	49 (34%)	140 (70%)	145 (72.1%)	298 (50.4%)	79 (39.5%)
1	24 (16.9%)	44 (22%)	34 (17%)	133 (62.1%)	79 (39.5%)
2	12 (8.3%)	12 (6%)	10 (5%)	73 (12.4%)	24 (12%)
3	13 (9%)	4 (2%)	12 (6%)	87 (14.7%)	18 (9%)
Urine Bilirubin					
0	137 (95.1%)	190 (95%)	190 (94.5%)	549 (94.9%)	194 (97%)
1	4 (2.8%)	10 (5%)	11 (5.5%)	28 (4.7%)	4 (2%)
2	2 (1.4%)	0	0	7 (1.2%)	2 (1%)
3	1 (0.7%)	0	0	7 (1.2%)	0
Urine Glucose					
0	86 (59.7%)	179 (89.5%)	183 (91%)	509 (86.1%)	185 (92.5%)
1	8 (5.6%)	12 (6%)	8 (4%)	52 (8.8%)	9 (4.5%)
2	1 (0.7%)	3 (1.5%)	2 (1%)	13 (2.2%)	3 (1.5%)
3	3 (2.1%)	6 (3%)	8 (4%)	17 (2.9%)	3 (1.5%)
Urine Ketone					
0	86 (59.7%)	184 (92%)	176 (87.6%)	523 (88.5%)	156 (78%)
1	10 (7%)	14 (7%)	24 (12%)	58 (9.8%)	38 (19%)
2	0	0	0	6 (1%)	5 (2.5%)
3	2 (1.4%)	2 (1%)	1 (0.5%)	4 (0.7%)	1 (0.5%)
Urine Protein					<i>p</i> < 0.0001
0	75 (52.1%)	124 (62%)	118 (58.7%)	279 (47.2%)	156 (78%)
Trace	16 (11.1%)	25 (12.5%)	25 (12.4%)	50 (8.5%)	15 (7.5%)
1	15 (10.4%)	19 (9.5%)	27 (13.4%)	91 (15.4%)	9 (4.5%)
2	24 (16.7%)	19 (9.5%)	23 (11.4%)	97 (16.4%)	12 (6%)
3	14 (9.7%)	13 (6.5%)	8 (4%)	74 (12.5%)	8 (4%)
Urobilinogen					<i>p</i> = 0.001
0	2 (1.4%)	7 (3.5%)	6 (3%)	20 (3.4%)	1 (0.5%)
0.1	55 (38.2%)	62 (31%)	59 (29.4%)	200 (33.8%)	124 (62%)
0.2	55 (38.2%)	84 (42%)	83 (41.3%)	261 (44.2%)	53 (26.5%)
1	29 (20.1%)	47 (23.5%)	52 (25.9%)	106 (17.9%)	20 (10%)
2	3 (2.1%)	0	1 (0.5%)	2 (0.3%)	1 (0.5%)
4	0	0	0	2 (0.3%)	1 (0.5%)

**Table 4.** Cont.

	<b>Cystitis n = 144</b>	<b>Kidney Cancer n = 200</b>	<b>Prostate Cancer n = 201</b>	<b>Bladder Cancer n = 591</b>	<b>Uterus Cancer n = 200</b>
Urine occult Blood		<i>p</i> < 0.0001	<i>p</i> = 0.001		<i>p</i> = 0.016
0	52 (36.1%)	119 (59.5%)	112 (55.7%)	201 (34%)	85 (42.5%)
Trace	16 (11.1%)	23 (11.5%)	22 (10.9)	66 (11.2%)	26 (13%)
1	12 (8.3%)	25 (12.5%)	16 (8%)	53 (9%)	28 (14%)
2	19 (13.2%)	13 (6.5%)	21 (10.4%)	73 (12.4%)	29 (14.5%)
3	45 (31.3%)	20 (10%)	30 (14.9%)	198 (33.5%)	32 (16%)

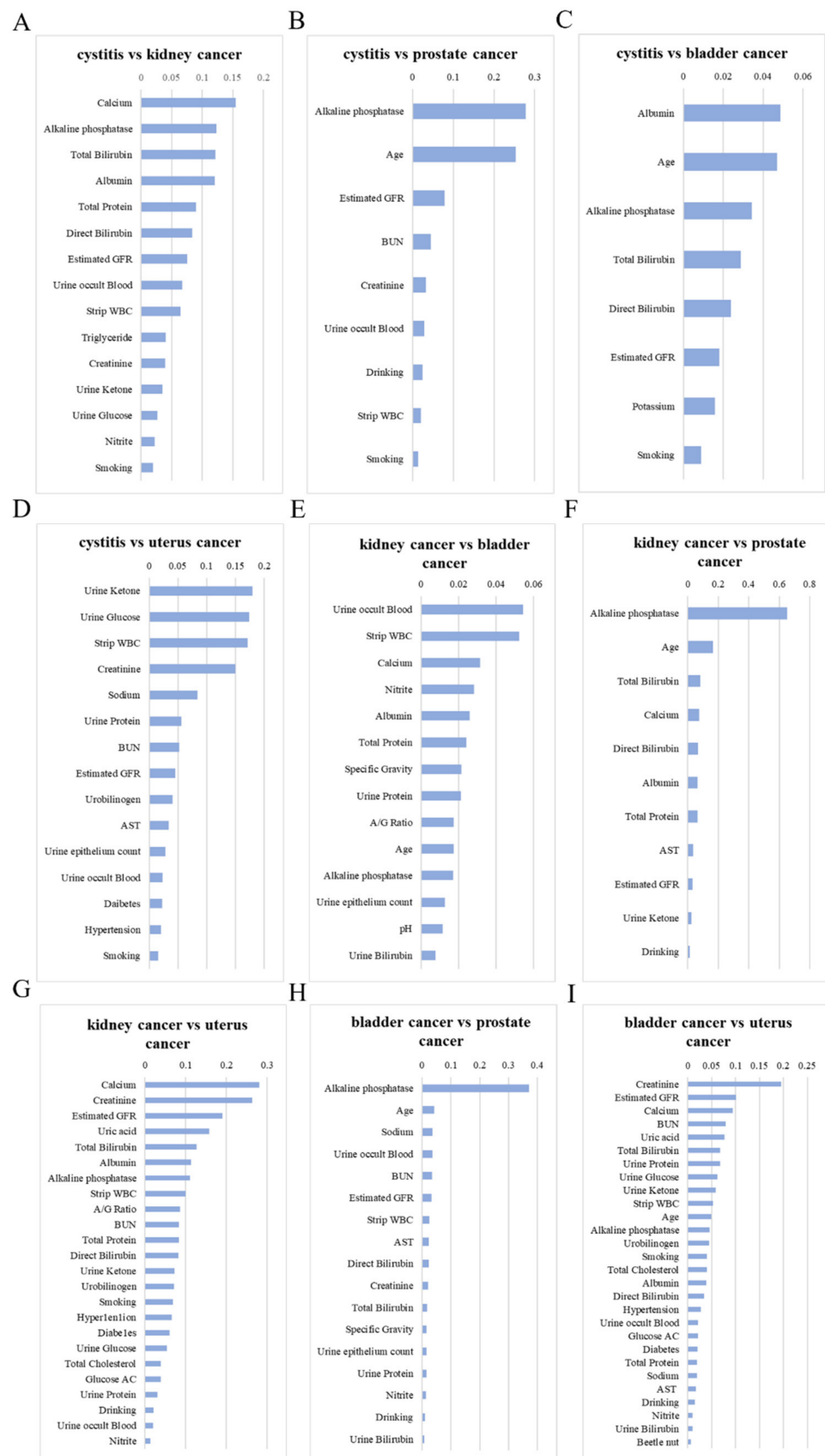
\* means *p* < 0.05, \*\* means *p* < 0.0001.

### 3.2. Feature Selection and Sampling Technique Experiment

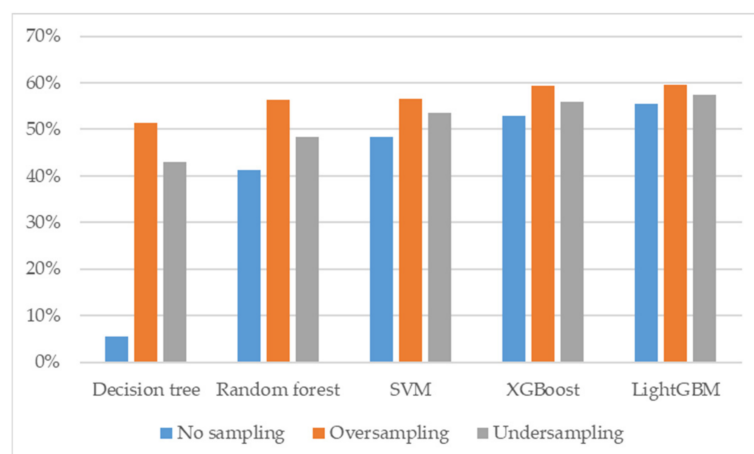
To reduce the noise in the dataset, we used InfoGain + Ranker to rank the features between groups (Figure 1). The top feature in each group was assembled as a set of selected features, including calcium, alkaline phosphate, albumin, urine ketone, urine occult blood, and creatinine (Table 5). We used the dataset to train and validate five models, including decision tree, random forest, SVM, XGBoost, and lightGBM. However, the evaluation parameter may not reflect the learning results, due to imbalanced data from bladder cancer compared to other groups. We used the python package named imbalanced-learn to solve the sample imbalance issue. Five models without any sampling techniques were trained and validated with the dataset. In differentiating patients with bladder cancer from patients with cystitis, the models received an accuracy 77.2–78.8%, precision 76.2–80.8%, f1 score 76.6–86.9%, sensitivity 77.7–95.8%, specificity 5–55.4%, and roc 0.592–0.729 (Table S1). After the oversampling technique was applied in the training and validating, the accuracy was adjusted to 73.4–78.8%, precision was adjusted to 74.9–81%, f1 score was adjusted to 75.6–81.4%, sensitivity was adjusted to 78–84.3%, and specificity was adjusted to 51.3–59.3% (Table S1 and Figure 2). An undersampling technique was also tested in our study. The accuracy was adjusted to 76.3–78.3%, precision was adjusted to 78.6–80.6%, f1 score was adjusted to 77.8–80.3%, sensitivity was adjusted to 79.0–83.9%, specificity was adjusted to 42.9–57.4%, and the roc was adjusted to 0.69–0.74 (Table S1). To further optimize our models, we conducted forward selection with the sampling technique in five different models.

**Table 5.** The top selected feature from each comparison group.

<b>Comparison Group</b>	<b>Top Selected Feature</b>
cystitis	kidney cancer prostate cancer bladder cancer uterus cancer
kidney cancer	Calcium ALP Albumin Urine Ketone
bladder cancer	Urine occult blood ALP Calcium
	prostate cancer uterus cancer
	ALP Creatinine



**Figure 1.** The feature selection results of (A) cystitis vs. kidney cancer, (B) cystitis vs. prostate cancer, (C) cystitis vs. bladder cancer, (D) cystitis vs. uterus cancer, (E) kidney cancer vs. bladder cancer, (F) kidney cancer vs. prostate cancer, (G) kidney cancer vs. uterus cancer, (H) bladder cancer vs. prostate cancer, and (I) bladder cancer vs. uterus cancer.



**Figure 2.** The specificity comparison of sampling techniques in five different models.

### 3.3. Model Evaluation and Comparison

The forward selection method is illustrated in Figure 3. The features from forward selection may be different, due to the models and the classes in the dataset. In the forward selection experiment, we focused on discriminating the patients with cystitis and patients with bladder cancer. In the results of the decision tree classifier, the features including ALT, AST, potassium, sodium, specific gravity, strip WBC, total protein, triglyceride, urine epithelium count, and uric acid were further selected. The decision tree classifier was trained and validated with features from WEKA and forward selection. The model received an accuracy of 76.2%, a precision of 77.9%, a f1 score of 74.6%, a sensitivity of 73.2%, a specificity of 78.1%, and an AUC of 0.77 in differentiating patients with bladder cancer from patients with cystitis (Table 6). In the results of the random forest classifier, the feature including ALT was selected. The random forest classifier was trained and validated with features from WEKA and forward selection. The model received an accuracy of 83.1%, a precision of 78.2%, a f1 score of 81.6%, a sensitivity of 85.5%, a specificity of 79.4%, and an AUC of 0.88 in discriminating patients with bladder cancer from patients with cystitis (Table 6). In the results of SVM, features including ALT, BUN, chloride, direct bilirubin, nitrite, and pH were further selected. The SVM was trained and validated with features from WEKA and forward selection. The model received an accuracy of 71.7%, a precision of 81.9%, a f1 score of 65.5%, a sensitivity of 55.7%, a specificity of 86.7%, and an AUC of 0.73 in identifying patients with bladder cancer and patients with cystitis (Table 6). In the results of XGBoost, features including ALT, AST, BUN, chloride, direct bilirubin, pH, potassium, sodium, total bilirubin, and total cholesterol were further selected (Table 6). The XGBoost model was trained and validated with features from WEKA and forward selection. The model received an accuracy of 82.8%, a precision of 84.7%, a f1 score of 82.7%, a sensitivity of 81.4%, a specificity of 83.3%, and an AUC of 0.87 in discriminating patients with bladder cancer from patients with cystitis (Table 6). In the results of lightGBM, features including ALT, and diabetes were further selected. The lightGBM model was trained with features from WEKA and forward selection. The model received an accuracy of 87.6%, a precision of 86.3%, a f1 score of 87.7%, a sensitivity of 89.5%, a specificity of 85.5%, and an AUC of 0.93 in identifying patients with bladder cancer and patients with cystitis (Table 6).



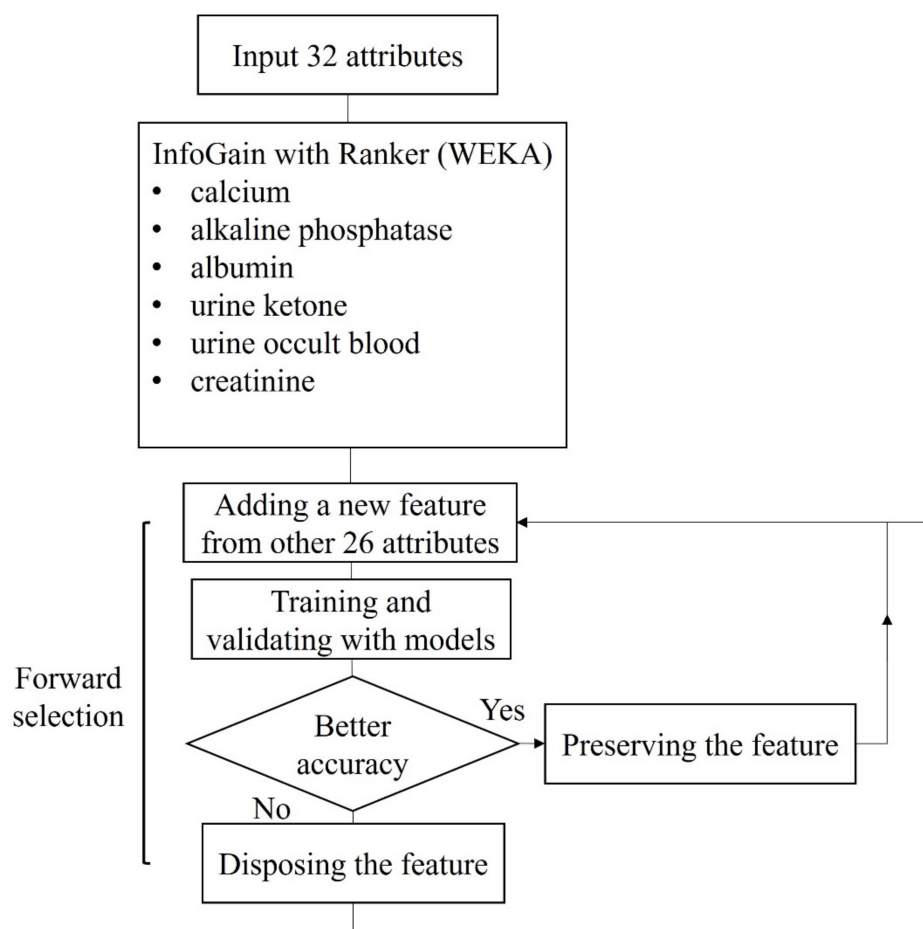


Figure 3. The block diagram of the forward selection method.

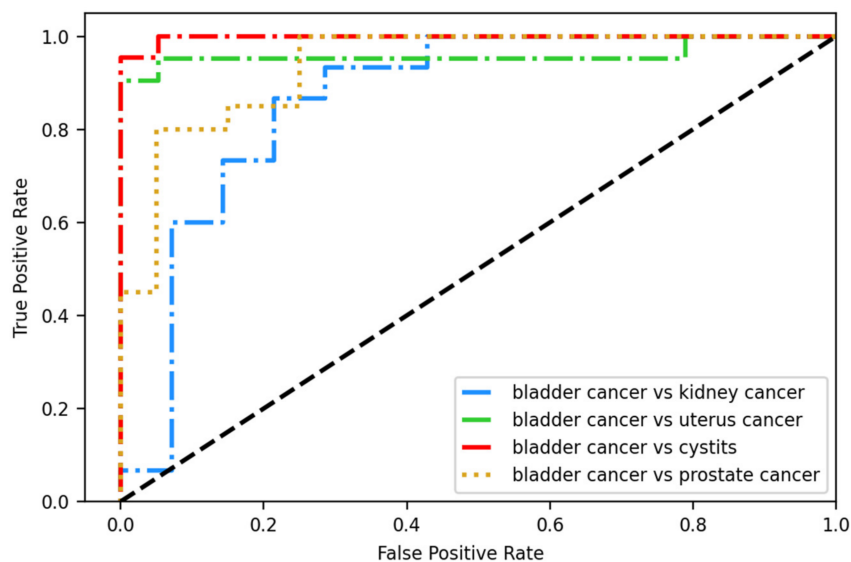
Table 6. Predictive performance of five models in discriminating patients with bladder cancer from patients with cystitis.

Models	Accuracy (95%CI)	Precision (95%CI)	f1 Score (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	AUC (95%CI)
decision tree	76.2% (71–81.5%)	77.9% (69.1–86.6%)	74.6% (69.8–79.5%)	73.2% (62.5–83.9%)	78.1% (64.9–91.3%)	0.775 (0.711–0.839)
random forest	83.1% (78.7–87.5%)	78.2% (71.5–85%)	81.6% (74.3–88.9%)	85.5% (76–94.9%)	79.4% (72–86.8%)	0.887 (0.826–0.947)
SVM	71.7% (63.5–80%)	81.9% (71.3–92.4%)	65.5% (51.9–79.1%)	55.7% (40.6–70.8%)	86.7% (78.9–94.5%)	0.736 (0.624–0.849)
XGBoost	82.8% (76.7–88.8%)	84.7% (74.5–94.9%)	82.7% (76.2–89.2%)	81.4% (75.1–87.7%)	83.3% (71–95.7%)	0.879 (0.819–0.939)
lightGBM	87.6% (81–94.1%)	86.3% (77.9–94.6%)	87.7% (81.8–93.5%)	89.5% (84.2–94.9%)	85.5% (75.2–95.7%)	0.932 (0.862–1.000)

The lightGBM with selected features received the highest score for accuracy, precision, f1 score, sensitivity, and AUC. Therefore, we further evaluated the model performance in differentiating bladder cancer from other cancers. The model received an accuracy of 84.8% to 86.9%, a precision of 83% to 87.1%, a f1 score of 84.5% to 87.7%, a sensitivity of 84.4% to 87.8%, a specificity of 82.9% to 86.7%, and an AUC of 0.88 to 0.92 (Table 7 and Figure 4).

**Table 7.** Predictive performance of the lightGBM model.

Groups	Accuracy (95%CI)	Precision (95%CI)	f1 Score (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	AUC (95%CI)
bladder cancer uterus cancer	86.9% (77.3–96.5%)	87.1% (76.5–97.7%)	87.3% (77.5–97%)	87.8% (77.3–98.3%)	86.7% (76.8–96.6%)	0.918 (0.849–0.988)
bladder cancer prostate cancer	84.8% (76–93.6%)	86.6% (76.8–96.4%)	85.1% (75.3–94.9%)	84.4% (71.9–96.9%)	85.1% (76.4–93.8%)	0.883 (0.823–0.942)
bladder cancer cystitis	87.6% (81–94.1%)	86.3% (77.9–94.6%)	87.7% (81.8–93.5%)	89.5% (84.2–94.9%)	85.5% (75.2–95.7%)	0.932 (0.862–1.001)
bladder cancer kidney cancer	84.5% (78.3–90.6%)	83% (73.2–92.9%)	84.5% (78.3–90.7%)	86.8% (79.9–93.7%)	82.9% (72.7–93.2%)	0.928 (0.88–0.977)
kidney cancer cystitis	86.2% (78.2–94.2%)	86.8% (78–95.6%)	86.9% (79–94.9%)	88% (76.5–99.6%)	84.2% (73–95.5%)	0.903 (0.854–0.952)
uterus cancer cystitis	83.8% (77.1–90.5%)	87% (76.8–97.3%)	83.7% (76.7–90.7%)	81.8% (71.2–92.5%)	86.5% (76.9–96.1%)	0.915 (0.834–0.995)
prostate cancer cystitis	87.6% (82.9–92.2%)	88.5% (80.3–96.6%)	87% (81.7–92.4%)	86.7% (77.5–95.8%)	88.4% (77.9–98.9%)	0.944 (0.903–0.986)
prostate cancer uterus cancer	84.1% (77–91.3%)	82.5% (68.2–96.7%)	82.3% (70.1–94.5%)	82.7% (70.5–94.9%)	84.7% (76.4–93.1%)	0.897 (0.837–0.957)
kidney cancer uterus cancer	86.2% (78.4–94%)	87.1% (76.5–97.7%)	86.8% (79–94.5%)	87.1% (79.5–94.6%)	86.2% (75.6–96.9%)	0.923 (0.858–0.988)
kidney cancer prostate cancer	84.5% (77.7–91.2%)	82.1% (72.7–91.6%)	84.9% (79.1–90.7%)	88.6% (82.5–94.6%)	81.3% (69.1–93.5%)	0.91 (0.849–0.972)

**Figure 4.** The receiver operating characteristic curve (ROC) plot of separating bladder cancer from cystitis or other cancers.

#### 4. Discussion

Machine learning has been significantly developed in the past decade. Many machine learning applications have been created with different types of data, including genomic data, transcriptomic data, proteomic data, image data, electronic health records (EHR), and clinical laboratory data [46–48]. However, the most intriguing question is whether machine learning can be applied to medical diagnosis [49]. Obermeyer et al. suggested that machine learning applied to clinical laboratory data can dramatically improve prognosis and diagnostic accuracy [50]. Moreover, compared to novel biomarkers, a decision-making

assist program based on clinical laboratory data can be considered a fast and cheap solution for improving the accuracy of diagnosis.

Herein, we utilized the clinical laboratory dataset coupled with machine learning algorithms to discriminate patients with bladder cancer from patients with cystitis. Missing values and imbalanced data were the two major challenges we encountered in this study. Missing values can be categorized as missing complete at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [51]. Several methods can solve this issue, such as collecting more samples, removing the subjects with missing values, filling with mean or median, and imputing missing values [52,53]. Multiple imputation (MI) is considered a good method to calculate missing values from existing data [54]. An elegant study from Hong et al. performed MI with models such as random forest and received a good accuracy [55]. However, in our experiment, MI did not receive applicable results (data not shown). We speculated that MI needs a larger sample size or strongly correlated features to obtain enough characteristics from the existing data. The skewed data distribution of one class over another is considered imbalanced data [56]. The imbalanced data causes classification problems during the training of machine learning algorithms [57]. Therefore, we used sampling techniques, including oversampling and undersampling, to reduce the error. The study performed by Mohammed et al. suggested that oversampling has a better performance with certain classifiers and evaluation metrics [58]. The oversampling method was applied to molecular description data by Chang et al., and reported that it could be used to reduce the overfitting problem [59]. However, oversampling has some disadvantages, such as sample overlapping, noise interference, and blindness of neighbor selection [60]. The main disadvantage of oversampling is that by making copies from existing data, overfitting is likely; in contrast, the main disadvantage of undersampling is the discarding of potentially useful data [61]. Instead of acquiring the highest performance from the models, our goal was to achieve an authentic performance from the models with our dataset. Furthermore, when faced with imbalanced data, it requires more than a one-step solution to improve the accuracy of the model [62]. Thus, in our experiment, we applied tools including undersampling techniques, feature selection, and improved models to increase the diagnostic accuracy in bladder cancer.

From our two-step feature selection, calcium, ALP, albumin, urine ketone, urine occult blood, and ALT were selected from the clinical laboratory data. Michel et al. reported that hypercalcemia was only observed in several patients with bladder cancer. Furthermore, the data shown that the hypercalcemia was caused by increasing levels from the tumor [63]. Moreover, Rosa et al. reported that an increasing level of calcium is common in various cancers, but rare in bladder cancer [64]. In addition, Huang et al. suggested that the elevation of calcium in blood may be considered as an indicator of bone metastasis in bladder cancer [65]. These studies suggested that calcium is a good feature for discriminating bladder cancer from other cancers. ALP has been considered a prognostic biomarker in patients with prostate cancer [66]. Therefore, ALP was selected as the top feature for prostate cancer versus cystitis or other cancers in our study. Furthermore, Braendengen et al. reported that an increased level of ALP in serum did not improve the accuracy of a bone scan used for evaluation precystectomy, which suggested a low correlation between ALP and bladder cancer [67]. These studies indicated that ALP can be used to identify prostate cancer from other cancers without interfering with the classification of bladder cancer. Albumin and globulin play an important role in immunity and inflammation; therefore, several studies have been proposed the ratio of albumin to globulin ratio as a biomarker in gastric cancer and lung cancer [68]. In addition, Quhal et al. reviewed the albumin to globulin ratio in 1096 patients with non-muscle-invasive bladder cancer and found that the ratios independently predicted the progression of disease [69]. Moreover, Tan et al. proposed that the ratio of albumin to ALP can be used as a prognostic biomarker in upper tract urothelial carcinoma [70]. Urine ketone is one of the routine urinalysis. Only a few studies reported an alteration of ketone in patients with bladder cancer in a metabolomics study [71]. However, ketone body in urine has been considered as a high correlation to diabetes [72]. Further-

more, the ketones in the blood and urine may indicate that the patients were suffering from diabetic ketoacidosis [73]. Moreover, a comprehensive systemic review suggested that diabetes mellitus was associated with bladder cancer [74]. These studies indicate that urine ketone is related to bladder cancer. Urine occult blood has been considered a screening indicator for bladder cancer [75]. Furthermore, a test for microhematuria found a strong correlation with bladder cancer in 46,842 patients [76]. However, urine occult blood can also be observed in other types of cancer, such as kidney cancer [77] or simply in a benign disease [78]. The lightGBM model we built in this study can discriminate bladder cancer from kidney cancer or cystitis with accuracies of 0.876 and 0.845. The ratio of AST and ALT was proposed as an indicator of liver function [79]. Recently, the ratio of AST and ALT has been discovered to have an association with bladder cancer [80]. Furthermore, Ha et al. suggested that the ratio of AST and ALT may further serve as a prognosis indicator in bladder cancer [81].

In this study, we trained and validated models with selected data from WEKA. Furthermore, a forward selection was performed with five different models, to optimize the model performance. Among the models, the lightGBM model had the highest performance, including an accuracy of 0.87, a precision of 0.86, a f1 score of 0.87, a sensitivity of 0.89, and an AUC of 0.93 in separating patients with bladder cancer from patients with cystitis (Table 6). Many studies have aimed at improving the diagnostic accuracy in bladder cancer. Wang et al. utilized machine learning algorithms to improve the tumor marker-based screening for multiple cancers and yielded a sensitivity of 0.81 and a specificity of 0.64 [82]. Shao et al. applied ultra-performance liquid chromatography coupled with time-of-flight mass spectrometry to acquire metabolites profiles in 152 samples from patients with bladder cancer and hernia; furthermore, the decision tree model embedded in this study obtained an accuracy of 76.6%, a sensitivity of 71.88%, and a specificity of 86.67% [83]. Wittmann et al. developed a random forest model with a set of metabolites selected based on statistical significance, metabolic pathway coverage, and fold difference from global metabolomics profiling of urine. Moreover, the model was tested in two independent cohorts and received an AUC of 0.81 to 0.78 [84]. Belugina et al. developed a non-invasive potentiometric multi-sensory system to perform urine analysis. In addition, various models were used in this study and received an accuracy of 76%, a sensitivity of 80%, and a specificity of 75% [85]. Kouznetsova et al. used two modeling methods, including multilayer perceptron (MLP) and stochastic gradient descent (SGD) with logistic regression loss function to discriminate bladder cancer patients with metabolite profiling. The best performing model was able to identify bladder cancer patients with an accuracy of 82.54% [43]. Compared to those studies, the model we performed in this study provided a better sensitivity and specificity. For future work, we aim to collect data from different cohorts. Moreover, we are eager to build a model that can differentiate bladder cancer from cystitis and other cancers in our next work.

## 5. Conclusions

In summary, we used two-step feature selection to select eight clinical laboratory tests and established a prediction model for bladder cancer with lightGBM. Furthermore, sample techniques were also used in our study and adjusted the imbalanced data. Our study indicated the potential of utilizing clinical laboratory data to detect cancer.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/diagnostics12010203/s1>, Table S1: Comparison of predictive performance with no sampling, oversampling, and undersampling.

**Author Contributions:** Conceptualization, I.-J.T. and C.-Y.L.; methodology, W.-C.S. and C.-L.L.; software, H.-D.W. and I.-J.T.; validation, I.-J.T. and W.-C.S.; investigation, I.-J.T. and W.-C.S.; data curation, I.-J.T. and W.-C.S.; writing—original draft preparation, I.-J.T.; writing—review and editing, H.-D.W. and C.-Y.L.; visualization, I.-J.T. and W.-C.S.; supervision, C.-Y.L. and H.-D.W.; project

administration, C.-Y.L. and H.-D.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study was approved by the institutional review board of the study hospital. The Institutional Review of Mackay Memorial Hospital approved the study protocol (20MMHIS200e (8 July 2020)). The study conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data was placed in the supplementary material and GitHub (<https://github.com/I-JUNG-TSAI/Predict-BC-Clinical-Laboratory-Data>, accessed on 14 June 2021).

**Acknowledgments:** The authors thank Shao-Chun Lin (National Yilan Senior High School, Yilan City, 260026 Taiwan) who validated the model performance.

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

## References

- Saginala, K.; Barsouk, A.; Aluru, J.S.; Rawla, P.; Padala, S.A.; Barsouk, A. Epidemiology of Bladder Cancer. *Med. Sci.* **2020**, *8*, 15. [[CrossRef](#)] [[PubMed](#)]
- Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [[CrossRef](#)]
- Burger, M.; Catto, J.W.F.; Dalbagni, G.; Grossman, H.B.; Herr, H.; Karakiewicz, P.; Kassouf, W.; Kiemenev, L.A.; La Vecchia, C.; Shariat, S.; et al. Epidemiology and Risk Factors of Urothelial Bladder Cancer. *Eur. Urol.* **2013**, *63*, 234–241. [[CrossRef](#)]
- Zhu, C.-Z.; Ting, H.-N.; Ng, K.-H.; Ong, T.-A. A review on the accuracy of bladder cancer detection methods. *J. Cancer* **2019**, *10*, 4038–4044. [[CrossRef](#)]
- Planz, B.; Jochims, E.; Deix, T.; Caspers, H.P.; Jakse, G.; Boecking, A. The role of urinary cytology for detection of bladder cancer. *Eur. J. Surg. Oncol.* **2005**, *31*, 304–308. [[CrossRef](#)] [[PubMed](#)]
- Hindmarsh, J.T.; Lyon, A.W. Strategies to promote rational clinical chemistry test utilization. *Clin. Biochem.* **1996**, *29*, 291–299. [[CrossRef](#)]
- Huang, X.-J.; Choi, Y.-K.; Im, H.-S.; Yarimaga, O.; Yoon, E.; Kim, H.-S. Aspartate Aminotransferase (AST/GOT) and Alanine Aminotransferase (ALT/GPT) Detection Techniques. *Sensors* **2006**, *6*, 756–782. [[CrossRef](#)]
- Sharma, U.; Pal, D.; Prasad, R. Alkaline phosphatase: An overview. *Indian J. Clin. Biochem.* **2014**, *29*, 269–278. [[CrossRef](#)] [[PubMed](#)]
- Epstein, E.; Kiechle, F.L.; Artiss, J.D.; Zak, B. The clinical use of alkaline phosphatase enzymes. *Clin. Lab. Med.* **1986**, *6*, 491–505. [[CrossRef](#)]
- Beddhu, S.; Ma, X.; Baird, B.; Cheung, A.K.; Greene, T. Serum alkaline phosphatase and mortality in African Americans with chronic kidney disease. *Clin. J. Am. Soc. Nephrol.* **2009**, *4*, 1805–1810. [[CrossRef](#)] [[PubMed](#)]
- Kendall, M.J.; Cockel, R.; Becker, J.; Hawkins, C.F. Raised serum alkaline phosphatase in rheumatoid disease. An index of liver dysfunction? *Ann. Rheum. Dis.* **1970**, *29*, 537–540. [[CrossRef](#)] [[PubMed](#)]
- Van Hoof, V.O.; Van Oosterom, A.T.; Lepoutre, L.G.; De Broe, M.E. Alkaline phosphatase isoenzyme patterns in malignant disease. *Clin. Chem.* **1992**, *38*, 2546–2551. [[CrossRef](#)] [[PubMed](#)]
- Wymenga, L.F.; Boomsma, J.H.; Groenier, K.; Piers, D.A.; Mensink, H.J. Routine bone scans in patients with prostate cancer related to serum prostate-specific antigen and alkaline phosphatase. *BJU Int.* **2001**, *88*, 226–230. [[CrossRef](#)] [[PubMed](#)]
- Blaustein, M.P. Sodium ions, calcium ions, blood pressure regulation, and hypertension: A reassessment and a hypothesis. *Am J. Physiol.* **1977**, *232*, C165–C173. [[CrossRef](#)]
- Bou-Abboud, E.; Nattel, S. Relative role of alkalosis and sodium ions in reversal of class I antiarrhythmic drug-induced sodium channel blockade by sodium bicarbonate. *Circulation* **1996**, *94*, 1954–1961. [[CrossRef](#)]
- Chovancova, B.; Liskova, V.; Babula, P.; Krizanova, O. Role of Sodium/Calcium Exchangers in Tumors. *Biomolecules* **2020**, *10*, 1257. [[CrossRef](#)]
- Waxman, S.G. Mechanisms of Disease: Sodium channels and neuroprotection in multiple sclerosis—current status. *Nat. Clin. Pract. Neurol.* **2008**, *4*, 159–169. [[CrossRef](#)]
- Rothschild, M.A.; Oratz, M.; Schreiber, S.S. ALBUMIN SYNTHESIS++Supported in part by the U.S. Public Health Service Grants AA 00959 and HL 09562. In *Albumin: Structure, Function and Uses*; Rosenoer, V.M., Oratz, M., Rothschild, M.A., Eds.; Pergamon: Oxford, UK, 1977; pp. 227–253.
- Oettl, K.; Stadlbauer, V.; Petter, F.; Greilberger, J.; Putz-Bankuti, C.; Hallström, S.; Lackner, C.; Stauber, R.E. Oxidative damage of albumin in advanced liver disease. *Biochim. Biophys. Acta* **2008**, *1782*, 469–473. [[CrossRef](#)]
- Nelson, J.J.; Liao, D.; Sharrett, A.R.; Folsom, A.R.; Chambless, L.E.; Shahar, E.; Szklo, M.; Eckfeldt, J.; Heiss, G. Serum albumin level as a predictor of incident coronary heart disease: The Atherosclerosis Risk in Communities (ARIC) study. *Am. J. Epidemiol.* **2000**, *151*, 468–477. [[CrossRef](#)]



21. Chen, Y.-H.; Feng, S.-Y.; Cai, W.-M.; Chen, X.-F.; Huang, Z.-M. The Relationship between C-Reactive Protein/Albumin Ratio and Disease Activity in Patients with Inflammatory Bowel Disease. *Gastroenterol. Res. Pract.* **2020**, *2020*, 3467419. [[CrossRef](#)]
22. Martin, H. Laboratory measurement of urine albumin and urine total protein in screening for proteinuria in chronic kidney disease. *Clin. Biochem. Rev.* **2011**, *32*, 97–102. [[PubMed](#)]
23. Borch-Johnsen, K.; Feldt-Rasmussen, B.; Strandgaard, S.; Schroll, M.; Jensen, J.S. Urinary Albumin Excretion. *Arterioscler. Thromb. Vasc. Biol.* **1999**, *19*, 1992–1997. [[CrossRef](#)] [[PubMed](#)]
24. Feher, J. (Ed.) 7.4-Tubular Reabsorption and Secretion. In *Quantitative Human Physiology*, 2nd ed.; Academic Press: Boston, MA, USA, 2017; pp. 719–729.
25. Washington, I.M.; Van Hoosier, G. Chapter 3-Clinical Biochemistry and Hematology. In *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents*; Suckow, M.A., Stevens, K.A., Wilson, R.P., Eds.; Academic Press: Boston, MA, USA, 2012; pp. 57–116.
26. Uchino, S. Creatinine. *Curr. Opin. Crit. Care* **2010**, *16*, 562–567. [[CrossRef](#)] [[PubMed](#)]
27. Echeverry, G.; Hortin, G.L.; Rai, A.J. Introduction to Urinalysis: Historical Perspectives and Clinical Application. In *The Urinary Proteome: Methods and Protocols*; Rai, A.J., Ed.; Humana Press: Totowa, NJ, USA, 2010; pp. 1–12.
28. Simerville, J.A.; Maxted, W.C.; Pahira, J.J. Urinalysis: A comprehensive review. *Am. Fam. Physician* **2005**, *71*, 1153–1162. [[PubMed](#)]
29. Lillian, A.; Mundt, K.S. Chemical Analysis of Urine. In *Graff's Textbook of Routine Urinalysis and Body Fluids*; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2010; Volume 1, pp. 35–53.
30. Cavanaugh, C.; Perazella, M.A. Urine Sediment Examination in the Diagnosis and Management of Kidney Disease: Core Curriculum 2019. *Am. J. Kidney Dis.* **2019**, *73*, 258–272. [[CrossRef](#)]
31. Ismail, A.A. When laboratory tests can mislead even when they appear plausible. *Clin. Med.* **2017**, *17*, 329–332. [[CrossRef](#)]
32. Haymond, S.; McCudden, C. Rise of the Machines: Artificial Intelligence and the Clinical Laboratory. *J. Appl. Lab. Med.* **2021**, *6*, 1640–1654. [[CrossRef](#)]
33. U.S. National Library of Medicine. *Machine Learning-MeSH*; U.S. National Library of Medicine: Bethesda, MD, USA, 2016.
34. Mahesh, B. Machine Learning Algorithms-A Review. *Int. J. Sci. Res. (IJSR)* **2020**, *9*, 381–386.
35. Banerjee, M.; Reynolds, E.; Andersson, H.B.; Nallamothu, B.K. Tree-Based Analysis. *Circ. Cardiovasc. Qual. Outcomes* **2019**, *12*, e004879. [[CrossRef](#)]
36. Chang, W.; Liu, Y.; Xiao, Y.; Yuan, X.; Xu, X.; Zhang, S.; Zhou, S. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics* **2019**, *9*, 178. [[CrossRef](#)]
37. Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Modeling* **2019**, *2019*, 4150–4158. [[CrossRef](#)]
38. Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Med. Inform. Decis. Mak.* **2010**, *10*, 16. [[CrossRef](#)] [[PubMed](#)]
39. Cai, Z.; Xu, D.; Zhang, Q.; Zhang, J.; Ngai, S.M.; Shao, J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol. Biosyst.* **2015**, *11*, 791–800. [[CrossRef](#)] [[PubMed](#)]
40. Gould, M.K.; Huang, B.Z.; Tammemagi, M.C.; Kinar, Y.; Shiff, R. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. *Am. J. Respir. Crit. Care Med.* **2021**, *204*, 445–453. [[CrossRef](#)]
41. Ibrahim Obaid, O.; Mohammed, M.; Abd Ghani, M.K.; Mostafa, S.; Al-Dhief, F. Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer. *Int. J. Eng. Technol.* **2018**, *7*, 160–166. [[CrossRef](#)]
42. Garapati, S.S.; Hadjiiski, L.; Cha, K.H.; Chan, H.P.; Caoili, E.M.; Cohan, R.H.; Weizer, A.; Alva, A.; Paramagul, C.; Wei, J.; et al. Urinary bladder cancer staging in CT urography using machine learning. *Med. Phys.* **2017**, *44*, 5814–5823. [[CrossRef](#)] [[PubMed](#)]
43. Kouznetsova, V.L.; Kim, E.; Romm, E.L.; Zhu, A.; Tsigelny, I.F. Recognition of early and late stages of bladder cancer using metabolites and machine learning. *Metabolomics* **2019**, *15*, 94. [[CrossRef](#)]
44. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
45. Tsai, K.L.; Chang, C.C.; Chang, Y.S.; Lu, Y.Y.; Tsai, I.J.; Chen, J.H.; Lin, S.H.; Tai, C.C.; Lin, Y.F.; Chang, H.W.; et al. Isotypes of autoantibodies against novel differential 4-hydroxy-2-nonenal-modified peptide adducts in serum is associated with rheumatoid arthritis in Taiwanese women. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 49. [[CrossRef](#)]
46. Liu, Y.; Bai, F.; Tang, Z.; Liu, N.; Liu, Q. Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation using atrial samples from patients with valvular heart disease. *BMC Cardiovasc. Disord.* **2021**, *21*, 52. [[CrossRef](#)]
47. Wong, J.; Murray Horwitz, M.; Zhou, L.; Toh, S. Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data. *Curr. Epidemiol. Rep.* **2018**, *5*, 331–342. [[CrossRef](#)]
48. Beam, A.L.; Kohane, I.S. Big Data and Machine Learning in Health Care. *JAMA* **2018**, *319*, 1317–1318. [[CrossRef](#)]
49. Gunčar, G.; Kukar, M.; Notar, M.; Brvar, M.; Černelč, P.; Notar, M.; Notar, M. An application of machine learning to haematological diagnosis. *Sci. Rep.* **2018**, *8*, 411. [[CrossRef](#)]
50. Obermeyer, Z.; Emanuel, E.J. Predicting the Future-Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219. [[CrossRef](#)]
51. Pedersen, A.B.; Mikkelsen, E.M.; Cronin-Fenton, D.; Kristensen, N.R.; Pham, T.M.; Pedersen, L.; Petersen, I. Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.* **2017**, *9*, 157–166. [[CrossRef](#)]

52. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 663. [[CrossRef](#)] [[PubMed](#)]
53. Royston, P. Multiple Imputation of Missing Values. *Stata J.* **2004**, *4*, 227–241. [[CrossRef](#)]
54. Patrician, P.A. Multiple imputation for missing data. *Res. Nurs. Health* **2002**, *25*, 76–84. [[CrossRef](#)] [[PubMed](#)]
55. Hong, S.; Lynn, H.S. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Med. Res. Methodol.* **2020**, *20*, 199. [[CrossRef](#)] [[PubMed](#)]
56. Kaur, H.; Pannu, H.S.; Malhi, A.K. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.* **2019**, *52*, 79. [[CrossRef](#)]
57. Wu, J.; Roy, J.; Stewart, W.F. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Med. Care.* **2010**, *48*, S106–S113. [[CrossRef](#)] [[PubMed](#)]
58. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 243–248.
59. Chang, C.Y.; Hsu, M.T.; Esposito, E.X.; Tseng, Y.J. Oversampling to overcome overfitting: Exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. *J. Chem. Inf. Model.* **2013**, *53*, 958–971. [[CrossRef](#)]
60. Jiang, Z.; Pan, T.; Zhang, C.; Yang, J. A New Oversampling Method Based on the Classification Contribution Degree. *Symmetry* **2021**, *13*, 194. [[CrossRef](#)]
61. Ganganwar, V. An overview of classification algorithms for imbalanced datasets. *Int. J. Emerg. Technol. Adv. Eng.* **2012**, *2*, 42–47.
62. Peng, Z.; Yan, F.; Li, X. Comparison of the Different Sampling Techniques for Imbalanced Classification Problems in Machine Learning. In Proceedings of the 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Qiqihar, China, 28–29 April 2019; pp. 431–434.
63. Michel, F.; Gattegno, B.; Meyrier, A.; Paillard, F.; Roland, J.; Scetbon, V.; Thibault, P. Paraneoplastic Hypercalcemia Associated with Bladder Carcinoma: Report of 2 Cases. *J. Urol.* **1984**, *131*, 753–755. [[CrossRef](#)]
64. La Rosa, A.H.; Ali, A.; Swain, S.; Manoharan, M. Resolution of hypercalcemia of malignancy following radical cystectomy in a patient with paraneoplastic syndrome associated with urothelial carcinoma of the bladder. *Urol. Ann.* **2015**, *7*, 86–87. [[CrossRef](#)]
65. Huang, P.; Lan, M.; Peng, A.F.; Yu, Q.F.; Chen, W.Z.; Liu, Z.L.; Liu, J.M.; Huang, S.H. Serum calcium, alkaline phosphatase and hemoglobin as risk factors for bone metastases in bladder cancer. *PLoS ONE* **2017**, *12*, e0183835. [[CrossRef](#)]
66. Li, D.; Lv, H.; Hao, X.; Hu, B.; Song, Y. Prognostic value of serum alkaline phosphatase in the survival of prostate cancer: Evidence from a meta-analysis. *Cancer Manag. Res.* **2018**, *10*, 3125–3139. [[CrossRef](#)] [[PubMed](#)]
67. Braendengen, M.; Winderen, M.; Fosså, S.D. Clinical significance of routine pre-cystectomy bone scans in patients with muscle-invasive bladder cancer. *Br. J. Urol.* **1996**, *77*, 36–40. [[CrossRef](#)] [[PubMed](#)]
68. Mao, M.J.; Wei, X.L.; Sheng, H.; Wang, X.P.; Li, X.H.; Liu, Y.J.; Xing, S.; Huang, Q.; Dai, S.Q.; Liu, W.L. Clinical Significance of Preoperative Albumin and Globulin Ratio in Patients with Gastric Cancer Undergoing Treatment. *Biomed. Res. Int.* **2017**, *2017*, 3083267. [[CrossRef](#)]
69. Qahal, F.; Pradere, B.; Laukhtina, E.; Sari Motlagh, R.; Mostafaei, H.; Mori, K.; Schuetfort, V.M.; Karakiewicz, P.I.; Rouprêt, M.; Enikeev, D.; et al. Prognostic value of albumin to globulin ratio in non-muscle-invasive bladder cancer. *World J. Urol.* **2021**, *39*, 3345–3352. [[CrossRef](#)] [[PubMed](#)]
70. Tan, P.; Xie, N.; Ai, J.; Xu, H.; Xu, H.; Liu, L.; Yang, L.; Wei, Q. The prognostic significance of Albumin-to-Alkaline Phosphatase Ratio in upper tract urothelial carcinoma. *Sci. Rep.* **2018**, *8*, 12311. [[CrossRef](#)] [[PubMed](#)]
71. Pinto, J.; Carapito, Â.; Amaro, F.; Lima, A.R.; Carvalho-Maia, C.; Martins, M.C.; Jerónimo, C.; Henrique, R.; Bastos, M.L.; Guedes de Pinho, P. Discovery of Volatile Biomarkers for Bladder Cancer Detection and Staging through Urine Metabolomics. *Metabolites* **2021**, *11*, 199. [[CrossRef](#)] [[PubMed](#)]
72. Laffel, L. Ketone bodies: A review of physiology, pathophysiology and application of monitoring to diabetes. *Diabetes Metab. Res. Rev.* **1999**, *15*, 412–426. [[CrossRef](#)]
73. Misra, S.; Oliver, N.S. Utility of ketone measurement in the prevention, diagnosis and management of diabetic ketoacidosis. *Diabet. Med.* **2015**, *32*, 14–23. [[CrossRef](#)] [[PubMed](#)]
74. Xu, Y.; Huo, R.; Chen, X.; Yu, X. Diabetes mellitus and the risk of bladder cancer: A PRISMA-compliant meta-analysis of cohort studies. *Medicine* **2017**, *96*, e8588. [[CrossRef](#)] [[PubMed](#)]
75. Chen, H.I.; Liou, S.H.; Loh, C.H.; Uang, S.N.; Yu, Y.C.; Shih, T.S. Bladder cancer screening and monitoring of 4,4'-methylenebis(2-chloroaniline) exposure among workers in Taiwan. *Urology* **2005**, *66*, 305–310. [[CrossRef](#)] [[PubMed](#)]
76. Matulewicz, R.S.; DeLancey, J.O.; Pavey, E.; Schaeffer, E.M.; Popescu, O.; Meeks, J.J. Dipstick Urinalysis as a Test for Microhematuria and Occult Bladder Cancer. *Bladder Cancer* **2017**, *3*, 45–49. [[CrossRef](#)] [[PubMed](#)]
77. Ingelfinger, J.R. Hematuria in Adults. *N. Engl. J. Med.* **2021**, *385*, 153–163. [[CrossRef](#)]
78. Gomes, C.M.; Sánchez-Ortiz, R.F.; Harris, C.; Wein, A.J.; Rovner, E.S. Significance of hematuria in patients with interstitial cystitis: Review of radiographic and endoscopic findings. *Urology* **2001**, *57*, 262–265. [[CrossRef](#)]
79. Wu, J.; Chen, L.; Wang, Y.; Tan, W.; Huang, Z. Prognostic value of aspartate transaminase to alanine transaminase (De Ritis) ratio in solid tumors: A pooled analysis of 9400 patients. *Onco. Targets Ther.* **2019**, *12*, 5201–5213. [[CrossRef](#)] [[PubMed](#)]

80. Laukhtina, E.; Mostafaei, H.; D'Andrea, D.; Pradere, B.; Quhal, F.; Mori, K.; Miura, N.; Schuettfort, V.M.; Sari Motlagh, R.; Aydh, A.; et al. Association of De Ritis ratio with oncological outcomes in patients with non-muscle invasive bladder cancer (NMIBC). *World J. Urol.* **2021**, *39*, 1961–1968. [[CrossRef](#)]
81. Ha, Y.S.; Kim, S.W.; Chun, S.Y.; Chung, J.W.; Choi, S.H.; Lee, J.N.; Kim, B.S.; Kim, H.T.; Yoo, E.S.; Kwon, T.G.; et al. Association between De Ritis ratio (aspartate aminotransferase/alanine aminotransferase) and oncological outcomes in bladder cancer patients after radical cystectomy. *BMC Urol.* **2019**, *19*, 10. [[CrossRef](#)] [[PubMed](#)]
82. Wang, H.Y.; Chen, C.H.; Shi, S.; Chung, C.R.; Wen, Y.H.; Wu, M.H.; Lebowitz, M.S.; Zhou, J.; Lu, J.J. Improving Multi-Tumor Biomarker Health Check-up Tests with Machine Learning Algorithms. *Cancers* **2020**, *12*, 1442. [[CrossRef](#)] [[PubMed](#)]
83. Shao, C.H.; Chen, C.L.; Lin, J.Y.; Chen, C.J.; Fu, S.H.; Chen, Y.T.; Chang, Y.S.; Yu, J.S.; Tsui, K.H.; Juo, C.G.; et al. Metabolite marker discovery for the detection of bladder cancer by comparative metabolomics. *Oncotarget* **2017**, *8*, 38802–38810. [[CrossRef](#)] [[PubMed](#)]
84. Wittmann, B.M.; Stirdivant, S.M.; Mitchell, M.W.; Wulff, J.E.; McDunn, J.E.; Li, Z.; Dennis-Barrie, A.; Neri, B.P.; Milburn, M.V.; Lotan, Y.; et al. Bladder cancer biomarker discovery using global metabolomic profiling of urine. *PLoS ONE* **2014**, *9*, e115870. [[CrossRef](#)]
85. Belugina, R.; Karpushchenko, E.; Sleptsov, A.; Protoshchak, V.; Legin, A.; Kirsanov, D. Developing non-invasive bladder cancer screening methodology through potentiometric multisensor urine analysis. *Talanta* **2021**, *234*, 122696. [[CrossRef](#)]