Golestan University of Medical Sciences

Jorjani Biomedicine Journal

Online ISSN: 2645-3509

Effective diagnosis of bladder cancer: Leveraging bioinformatics and machine learning techniques

Nahid Nematy [1], Emadoddin Moudi [2], Masoud Arabfard [3]*

1. Pasteur institute of Iran, Tehran, Iran
2. Clinical Research Development Center, Shahid Beheshti Hospital, Babol University of Medical Sciences, Babol, Iran
3. Artificial Intelligence in Health Research Center, Biomedicine Technologies Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran

* Correspondence: Masoud Arabfard. Artificial Intelligence in Health Research Center, Biomedicine Technologies Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran. Tel: +989132649405; Email: arabfard@gmail.com

Abstract

Background: Bladder cancer (BC) is a life-threatening malignancy that can be successfully treated if diagnosed in its early stages. Machine learning techniques, by using large biological databases, are suggested as important approaches for identifying accurate diagnostic biomarkers. The present study aimed to introduce a simple and accurate model for the diagnosis of BC.

Methods: RNA-sequencing information of 412 primary bladder tumors versus 19 normal bladder tissues from The Cancer Genome Atlas were analyzed using the TCGAbiolinks R package to identify differentially expressed genes (DEGs). Gene ontology properties and the corresponding pathways of DEGs were investigated using the online ShinyGO tools. To develop a diagnostic model for BC, two binary classifier machine learning algorithms, C5.0 and CHAID, were employed in three subgroups of train, test, and validation sets using SPSS Modeler version 18.1. Their efficacy was evaluated using performance measures for binary classification.

Results: Most of the identified DEGs were associated with microtubule organization, coagulation, and myelination. Based on the constructed models, four important RNAs (Tubulin Polymerization-Promoting Protein: ENSG00000171368, Proteolipid Protein-1: ENSG00000123560, RP11-473E2: ENSG00000228877, and Coagulation Factor X: ENSG00000126218) were identified as important classifiers for diagnosis in both C5.0 and CHAID models. The CHAID model demonstrated superior performance in the testing dataset, achieving an accuracy of 98.75%, an F1-score of 99.36%, and an AUC of 99.4%.

Conclusion: According to the results, machine learning algorithms are beneficial for the diagnosis of BC and potentially useful for improving personalized medicine in BC patients. The developed model may serve as a non-invasive, data-driven tool to support early diagnosis and personalized treatment planning in clinical settings. Further evaluation using laboratory tests is suggested to validate the obtained results.

Article Type: Research Article

Article History

Received: 7 March 2025
Received in revised form: 3 May 2025
Accepted: 20 June 2025
Available online:
DOI: 10.29252/jorjanibiomedj.13.X.X

Keywords

Bioinformatic
Urinary Bladder Neoplasms
Machine Learning
Precision Medicine

Highlights

What is current knowledge?

- Bladder cancer is a life-threatening malignancy that can be successfully treated if diagnosed in early stages.
- Current diagnosis of bladder cancer relies on imaging, cystoscopy, and urinary cytology, which have limitations.
- Many bladder cancer patients are asymptomatic in early stages, complicating timely diagnosis.
- The Cancer Genome Atlas provides comprehensive gene expression and clinical information for cancer research.
- Machine learning algorithms can analyze high-dimensional transcriptomic data to identify important diagnostic biomarkers.

What is new here?

- This study developed simple and accurate machine learning models (C5.0 and CHAID) for diagnosing bladder cancer.
- Four important RNAs were identified as crucial classifiers for bladder cancer diagnosis.
- The CHAID model showed superior performance in testing data, achieving 98.75% accuracy and 99.4% AUC.
- Gene ontology analysis revealed that most identified differentially expressed genes (DEGs) are associated with mitochondrial organization, coagulation, and myelination.
- This study demonstrates the potential of machine learning algorithms in improving personalized medicine for bladder cancer patients.

Introduction

A large number of annual mortalities worldwide have been reported as a result of different malignancies (1). Bladder cancer (BC) is a well-known life-threatening tumor and based on its manifestations is mainly divided into ordered categories including non-muscle-invasive (Sub-grouped into low-grade papillary BC (Ta), carcinoma in situ (CIS), and high-grade T1 tumors), muscle-invasive subtypes, and metastatic BC (2). This classification is important since urologists must adopt various treatment strategies (3). Transurethral resection of bladder tumors along with intravesical instillation of Bacillus Calmette-Guérin (BCG) are the two main treatment approaches for low-grade Ta, CIS, and high-grade T1 non-muscle-invasive subtypes. Despite the risk of recurrence, these approaches are effective in several cases. However, radical cystectomy is preferred for patients with high-grade muscle-invasive BC. Indeed, the diagnostic time for BC and its progression are determining factors in the management of BC. In other words, patients diagnosed at early stages have a higher chance of successful treatment (3,4).

BC is often diagnosed accidentally through urinary cytology in patients presenting with hematuria, which is a common symptom caused by various pathological conditions such as urinary system stones, urinary tract infections, and malignancy (5). However, most cases are asymptomatic in the early stages and therefore lose the “golden time” for optimal management. To the best of our knowledge, BC lacks any specific biomarker for diagnosis or treatment monitoring. Most patients are identified through imaging, cystoscopy, and urinary cytology. However, these techniques have limitations such as low specificity and invasiveness. Although traditional methods are widely used in clinical practice, many have significant limitations, including insufficient sensitivity for early-stage tumors, invasiveness, and dependency on operator expertise. For instance, although urinary cytology is non-invasive, it often fails to detect low-grade tumors, and cystoscopy is expensive and uncomfortable for patients.

Multi-omics assessments are emerging as promising approaches for medical studies (6,7). It is well established that changes in gene expression occur at the onset of many physiological and pathological conditions. This means that comparing the transcriptome of patients with healthy individuals or conducting time-series analysis may help identify unique patterns useful for clinical applications. However, identifying such patterns, in addition to high costs, usually requires a sufficient sample size, advanced equipment, and technical expertise. Therefore, many researchers prefer not to use high-throughput techniques.

The use of reliable data repositories, such as The Cancer Genome Atlas (TCGA), is a valuable alternative to high-throughput laboratory approaches (8). TCGA comprises a comprehensive repository including gene expression data, mutation profiles, technical details, and clinical information for each sample. Depending on the research objective, these data allow the identification of exclusive features that, along with laboratory validation, can be used for scientific purposes. However, a major challenge is the high-dimensional transcriptomic data obtained from laboratory sequencing or genomic databases like TCGA, which often present researchers with large and complex matrices of genes and samples (9,10). Therefore, their interpretation to identify differentially expressed genes (DEGs) requires precise statistical software and machine learning algorithms.

Given the limitations of conventional diagnostic approaches and the challenges of interpreting high-dimensional data, there is an essential need for alternative diagnostic methods that are both precise and minimally invasive. In this context, machine learning (ML) models trained on transcriptome data offer a promising avenue for identifying reliable biomarkers and enhancing early detection of bladder cancer. Advanced analytical approaches, such as ML algorithms developed based on accurate historical data, clinical information, and high-throughput datasets, facilitate the identification of important biomarkers that may improve diagnostic accuracy and treatment outcomes (11,12). ML algorithms serve as valuable toolkits for the efficient and accurate diagnosis and therapeutic monitoring of diseases. In one ML technique-supervised learning-a large amount of data is used as input for a predefined target feature, training the algorithm to identify unique patterns that predict disease outcomes or detect accurate panels of diagnostic biomarkers (13).

Recent advancements in data mining, high-throughput repositories, and mathematical modeling, including ML algorithms, have provided clinicians with new perspectives for transforming personalized medicine (14,15). C5.0 and CHAID are two well-established machine learning classification algorithms used for diagnostic purposes. C5.0 is recognized for high accuracy, the ability to process large datasets, and its production of interpretable decision trees with minimal overfitting. CHAID, on the other hand, is effective in uncovering statistically significant relationships between variables, particularly in categorical data analysis. Therefore, both algorithms appear suitable for gene expression profiling and developing robust diagnostic models for bladder cancer. These approaches leverage multi-omics data to identify candidate biomarkers associated with enhanced diagnostic accuracy and treatment outcomes based on individual genetic profiles (16).

Objectives

In this study, regarding the importance of transcriptome analysis, we aimed to utilize machine learning algorithms to investigate candidate genes that are differentially expressed in bladder cancer patients. In clinical settings, this approach may assist physicians in personalized medicine.

Methods

Data source and pre-processing

The gene expression profiles of 412 primary tumor tissues of BC versus 19 solid normal bladder tissues were obtained from The Cancer Genome Atlas (TCGA) database with the specific project ID “TCGA-BLCA”. The inclusion criteria for selecting tumor samples were as follows: (1) Primary bladder tumor tissue samples with available RNA-sequencing data in FPKM format, and (2) complete clinical metadata including age, gender, tumor stage, and survival status. Normal samples included histologically confirmed solid normal bladder tissues from non-cancerous individuals within the TCGA-BLCA cohort. Samples were excluded if they were metastatic or recurrent tumor tissues. Other exclusion criteria included incomplete expression profiles, incomplete metadata, or missing clinical information.

Expression data analysis to identify differentially expressed genes (DEGs) and generate data matrices was executed using R software (R Foundation for Statistical Computing, Vienna, Austria). The dataset, named TCGA-BLCA, was obtained using the “GDCquery” function from the TCGAbiolinks R package in count format. After excluding samples with incomplete metadata or missing expression count data, missing values for categorical variables in the remaining samples were addressed using automatic imputation in SPSS Modeler. Normalization of expression data was performed to adjust for differences in sequencing depth and gene length using the fragments per kilobase of transcript per million (FPKM) method via the TCGAbiolinks package. These preprocessing steps were implemented to enhance sample comparability and improve the robustness of subsequent machine learning models. Basic clinical information, including gender, age at diagnosis, clinical stage, tumor grade, overall survival (OS) time, and survival status, was downloaded from the TCGA portal.

Evaluation of DEGs and candidate RNAs for BC

After identifying the DEGs, further investigations were performed regarding their corresponding proteins in terms of gene ontology (GO), including biological process, cellular component, and molecular function. In addition, pathway analysis to identify the most involved cellular pathways disrupted in BC was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG). For gene identification, Ensembl gene IDs were used through the online g:Profiler tool (https://biit.cs.ut.ee/gprofiler/page/citing) (17). Both GO enrichment and KEGG analyses were performed using ShinyGO (http://bioinformatics.sdstate.edu/go/), a graphical online bioinformatics tool developed at South Dakota State University (18).

Development and evaluation of ML models

Machine learning (ML) models were developed using a supervised learning approach in IBM SPSS Modeler 18.1 software. The data were randomly partitioned into three subgroups consisting of 70%, 20%, and 10% of the dataset. Diagnostic models for BC were constructed using the first 70% of the data as the training set and evaluated using the subsequent 20% as the testing set.

Feature selection was performed in three steps: Screening, ranking, and selection. The process began with screening expression data, where variables with low variance and non-informative features were eliminated. In the next step, ranking was conducted based on chi-square statistical results, and the top 150 genes were selected according to their cumulative contribution to model performance. The threshold of 150 genes was empirically set to balance dimensionality reduction with maintaining diagnostic accuracy.

Two widely used binary classification algorithms, C5.0 and CHAID, were employed to establish the diagnostic models. The rationale for selecting these algorithms was their demonstrated effectiveness in categorizing high-dimensional datasets, which is a key characteristic of transcriptomic data. The remaining 10% subgroup was utilized to validate the efficacy of the constructed models using key metrics for binary classification (19).

The C5.0 algorithm was developed with the default boosting option enabled, with a maximum of 10 boosting trials. Pruning severity was set at 75 to mitigate overfitting. For the CHAID model, a significance level of 0.05 was applied for both splitting and merging criteria. The minimum number of cases for parent and child nodes was set at 10 and 5, respectively. Missing values were handled using automatic imputation. These configurations were chosen to maintain a balance between model complexity and generalizability while ensuring interpretability.

Results

The study included gene expression profiles of 412 primary tumor tissues of BC versus 19 solid normal bladder tissues. The results of basic clinical information, including gender, primary diagnosis, age at diagnosis, history of prior malignancy, history of prior treatment, and survival status, are demonstrated in Table 1.

Table 1. Basic clinical and demographical information of bladder cancer patients

Feature	N (%)
Gender	
Male	303 (74.08)
Female	106 (25.92)
Primary diagnosis	
Transitional cell carcinoma	343 (83.86)
Papillary transitional cell carcinoma	66 (16.14)
Age at diagnosis (Years)	
34.38 to ≤ 45.5	8 (1.96)
45.51 to ≤ 56.63	46 (11.27)
56.64 to ≤ 67.75	137 (33.58)
67.76 to ≤ 78.87	143 (35.05)
78.88 to ≤ 90	74 (18.14)
Prior malignancy	
Yes	109 (26.65)
No	300 (73.35)
Prior treatment	
Yes	10 (2.44)
No	399 (97.56)
Survival status	
Alive	227 (55.64)
Death	181 (44.36)

Enrichment analysis and identifying biological pathways involved in BC

Enrichment analysis of DEGs is demonstrated in Figure 1. Based on the results, in terms of biological processes (Figure 1A), most of the identified genes are involved in myofibril assembly, striated muscle cell development, regulation of smooth muscle contraction, and cellular component assembly involved in morphogenesis. The cellular components of most DEGs include costamere, myofilament, striated muscle thin filaments, stress fibers, and contractile actin filament bundles (Figure 1B). In terms of molecular function, most of the identified DEGs contribute to tropomyosin binding, structural constituents of muscle, vinculin binding, and calcium-activated potassium channel activity (Figure 1C).

The results of biological pathways associated with the identified DEGs are demonstrated in Figure 2. According to the KEGG database, most DEGs are involved in the calcium signaling pathway, regulation of the actin cytoskeleton, and circadian entrainment.

Differentially Expressed Genes (DEGs) analysis and feature selection

A total of 32,765 Ensembl stable IDs were considered as input for developing the diagnostic model. Due to the high dimensionality of the input data, a feature selection algorithm was used. This algorithm consists of three main steps: Screening, ranking, and selecting. The overall workflow included removing unimportant inputs, sorting the remaining inputs based on their importance, and selecting the top informative features. Following the screening step, 8,729 DEGs were ranked, and the top 150 were selected for developing the models.

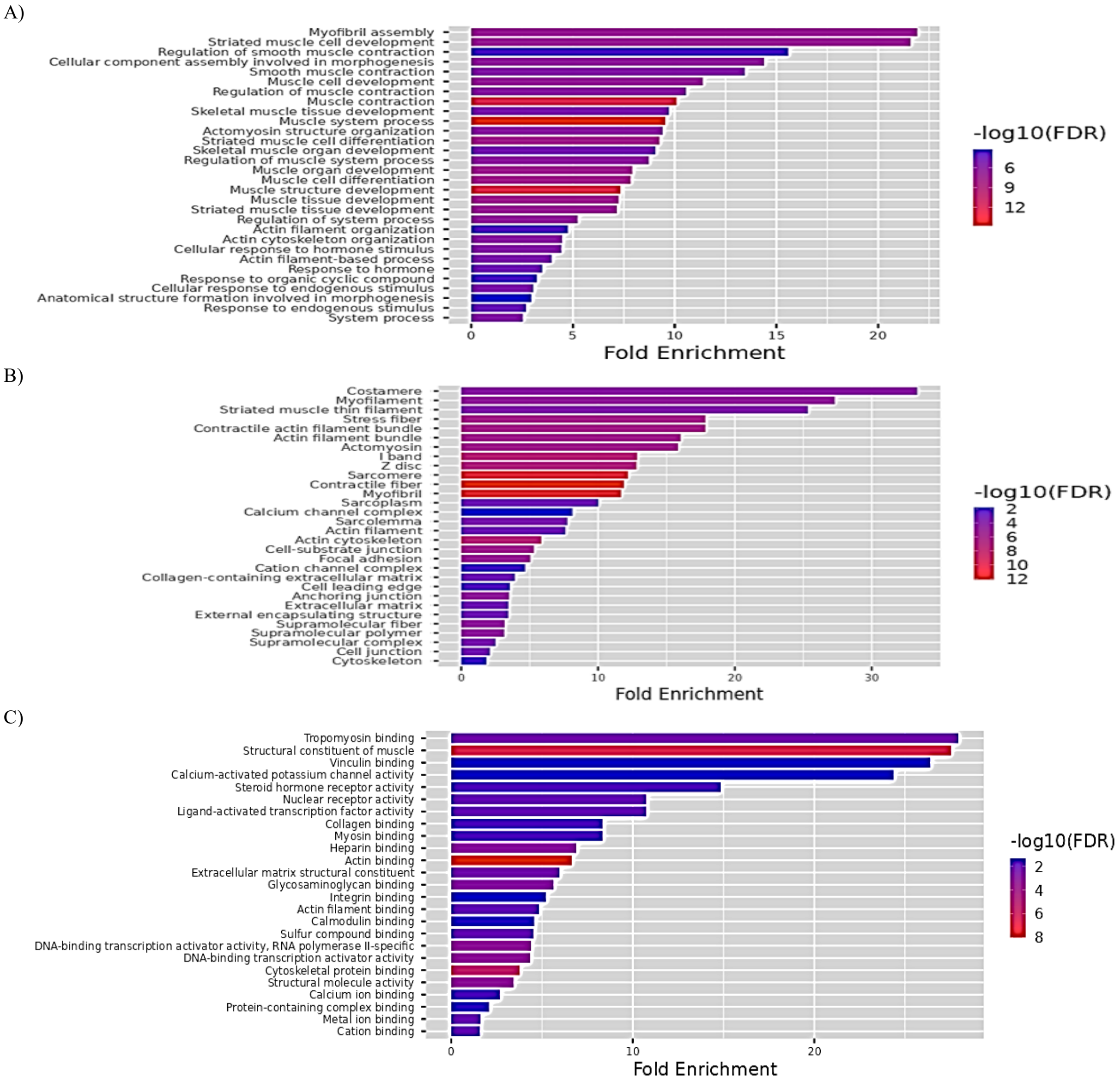


Figure 1. Gene ontology terms of DEGs in bladder cancer

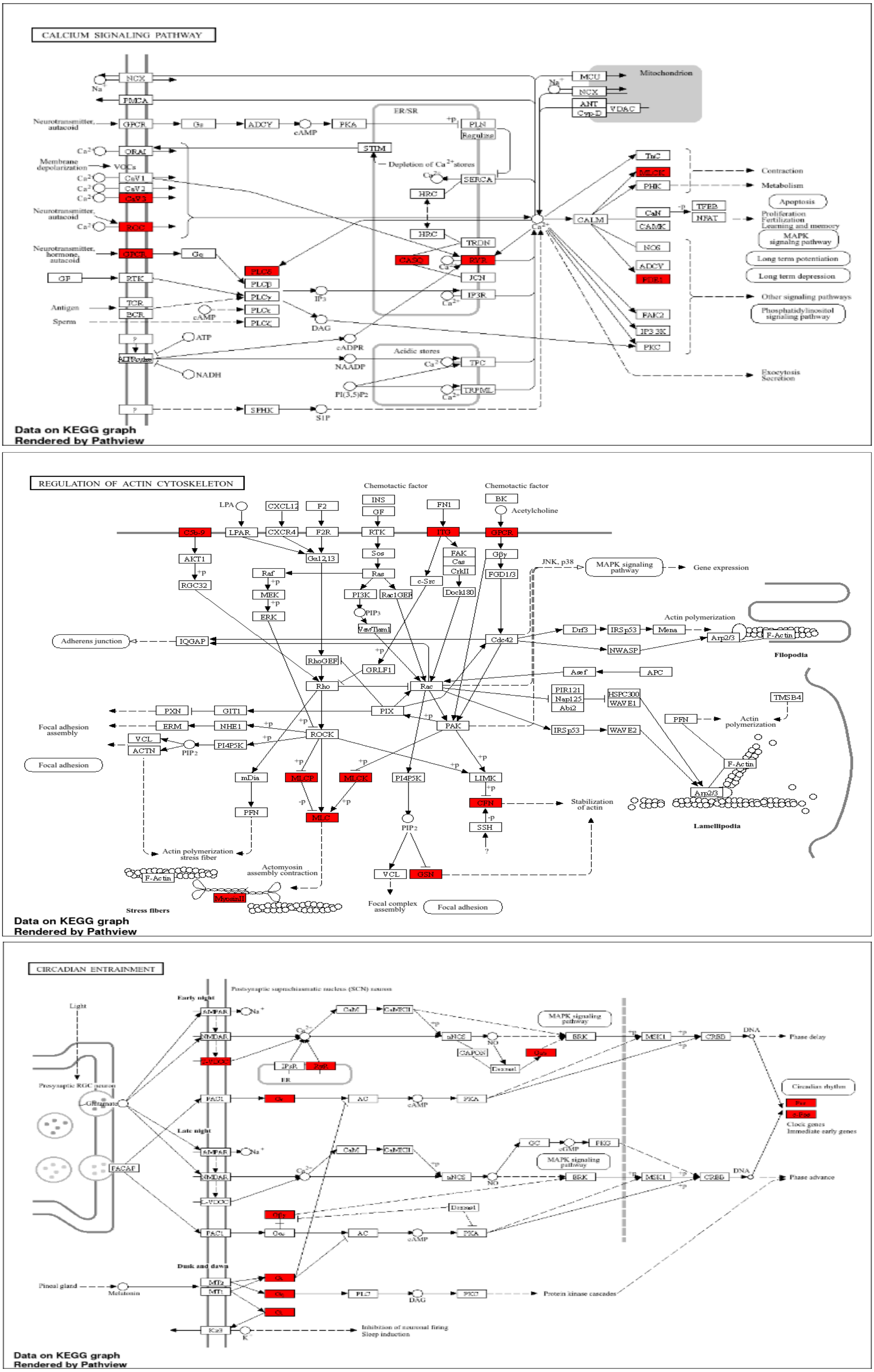


Figure 2. Biological pathways associated with DEGs in bladder cancer

C5.0 algorithm results

The C5.0 algorithm, commonly associated with decision trees, can be effectively employed for conducting classification and DEG-related analyses on RNA sequencing (RNA-seq) data. This model was developed using two important RNAs: ENSG00000171368 and ENSG00000126218 (Figure 3). The model can be interpreted as follows: If the expression level of ENSG00000171368 < 4.066, the model diagnoses the sample as primary tumor. Otherwise, the expression level of ENSG00000126218 is considered. Diagnosis of BC is possible when ENSG00000171368 < 4.066 or when ENSG00000171368 > 4.066 along with ENSG00000126218 < 1.0601. The confusion matrix for the developed C5.0 model is represented in Table 2.

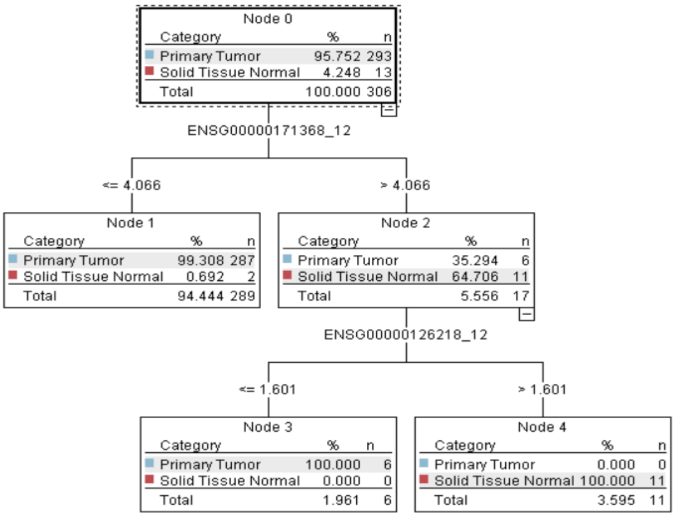


Figure 3. Schematic diagram of the C5.0 algorithm

CHAID algorithm results

Chi-squared Automatic Interaction Detector (CHAID) is a decision-tree algorithm particularly useful for analyzing RNA-seq data and performing classification tasks in bioinformatics. This methodology employs significance testing using chi-square statistics to establish relationships between the dependent variable and independent variables by creating appropriate data splits. The CHAID model was developed using two important RNAs: ENSG00000123560 and ENSG00000228877 (Figure 4). It can be interpreted as follows: If ENSG00000123560 < 0.473, the model diagnoses the sample as primary tumor. Otherwise, the expression level of ENSG00000228877 must be considered. Diagnosis of BC is possible when ENSG00000123560 < 0.473 or when ENSG00000123560 > 0.473 along with ENSG00000228877 < 0.264. The confusion matrix for the developed CHAID model is represented in Table 2.

The efficacy of the developed models for the diagnosis of BC was assessed using key metrics for binary classification. The evaluation metrics for training, testing, and validation datasets are shown in Table 3.

For the diagnosis of bladder cancer, it is essential to employ a method with high sensitivity to ensure detection at early stages and minimize false-negative results. High specificity is also important to reduce false positives and avoid unnecessary invasive interventions. Accuracy above 95% is generally considered ideal for clinical decision-making.

The observed metrics in both models suggest strong diagnostic potential, with CHAID showing particularly robust generalizability in the validation phase. In the training dataset, both C5.0 and CHAID models demonstrated good performance across all evaluation metrics. Although C5.0 demonstrated slightly higher sensitivity, accuracy, MCC, and no false negatives-indicating better diagnostic efficacy-the CHAID model performed better in testing and validation analyses, ultimately achieving 100% sensitivity and specificity with zero false results in the validation dataset.

A ROC curve is useful for interpreting and evaluating the diagnostic performance of models. The overall accuracy, sensitivity, and specificity of the C5.0 model, along with its corresponding ROC curve, are represented in Figure 5A. Based on the figure, the C5.0 model demonstrates strong performance with accuracy = 98.61%, sensitivity = 99.51%, and specificity = 78.94%, with only a small proportion of false results. The CHAID model achieved accuracy = 99.01%, sensitivity = 99.76%, and specificity = 84.21% (Figure 5B).

In general, both models performed well in the diagnosis of BC. Based on the models, four RNAs - ENSG00000171368, ENSG00000126218, ENSG00000123560, and ENSG00000228877 - were identified as important diagnostic classifiers for bladder cancer.

Table 2. The confusion matrix for all training, test, and validation data

Dataset	Tissue type	CHAID		C5		Total
		Primary tumor	Solid normal tissue	Primary tumor	Solid normal tissue	
Train	Primary tumor	292	1	293	0	293
	Solid normal tissue	2	11	2	11	13
Test	Primary tumor	78	0	76	2	78
	Solid normal tissue	1	1	1	1	2
Validation	Primary tumor	41	0	41	0	41
	Solid normal tissue	0	4	1	3	4

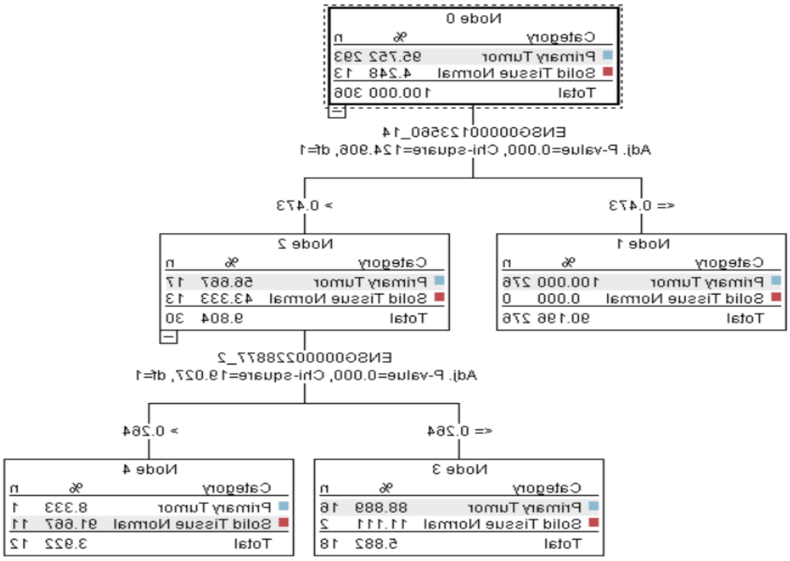
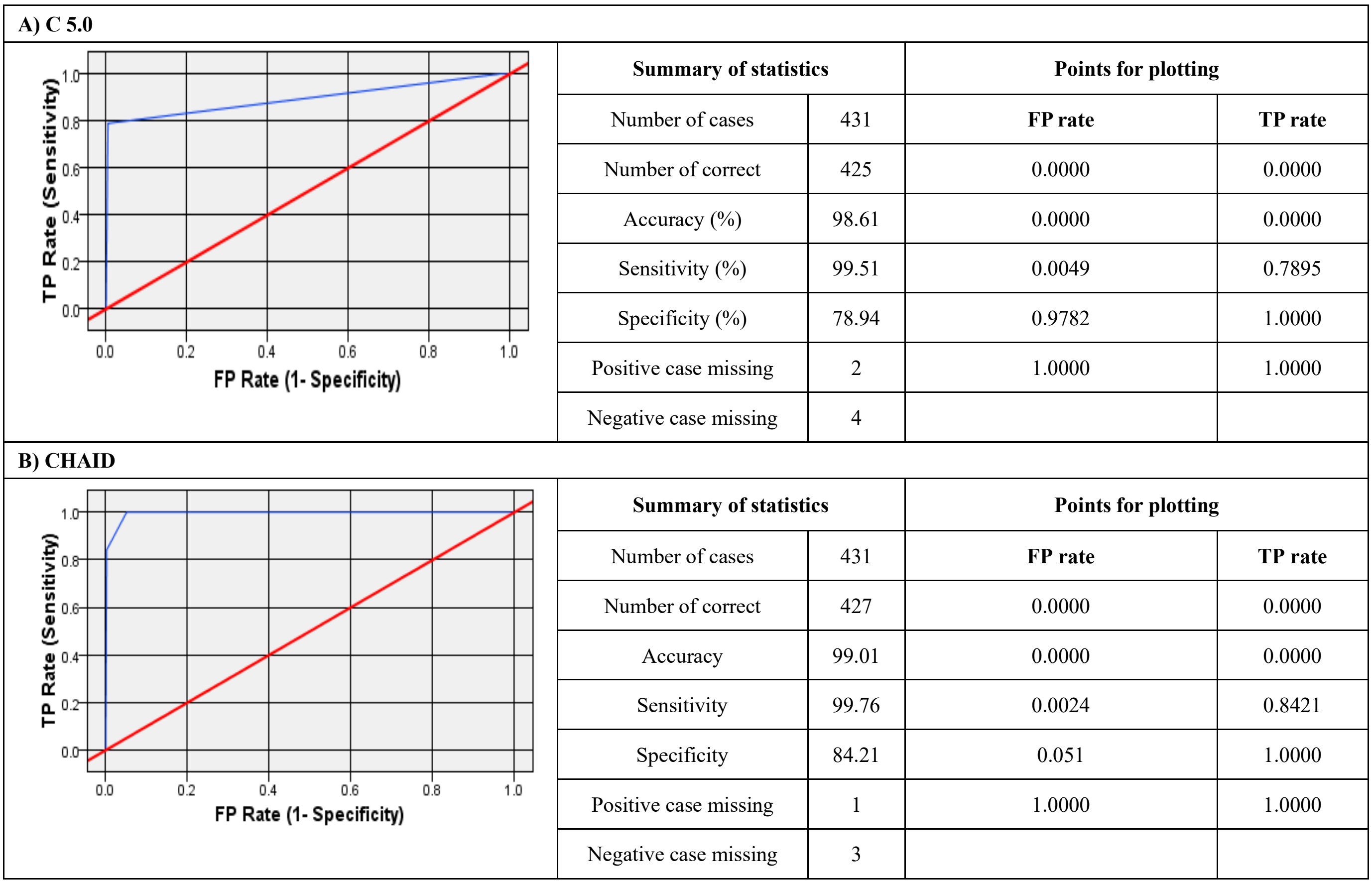


Figure 4. Schematic diagram for the CHAID algorithm

Table 3. The evaluation metrics for data mining models

Measure	CHAID	C 5.0
Training data		
Sensitivity	99.66	1.00
Specificity	84.62	84.62
False negative rate	0.34	0.00
False positive rate	15.38	15.38
Negative predictive value	91.67	1.00
Positive predictive value	99.32	99.32
Accuracy	99.02	99.35
F1 score	99.49	99.66
Matthew's correlation coefficient	87.57	91.67
Testing data		
Sensitivity	100.00	97.44
Specificity	50.00	50.00
False negative rate	0.00	2.56
False positive rate	50.00	50.00
Negative predictive value	100.00	33.33
Positive predictive value	98.73	98.70
Accuracy	98.75	96.25
F1 score	99.36	98.06
Matthew's correlation coefficient	70.26	38.98
Validation data		
Sensitivity	100.00	100.00
Specificity	100.00	75.00
False negative rate	0.00	0.00
False positive rate	0.00	0.25
Negative predictive value	100.00	100.00
Positive predictive value	100.00	97.62
Accuracy	100.00	97.78
F1 score	100.00	98.80
Matthew's correlation coefficient	100.00	85.57



A) C 5.0

Summary of statistics		Points for plotting	
		FP rate	**TP rate**
Number of cases	431		
Number of correct	425	0.0000	0.0000
Accuracy (%)	98.61	0.0000	0.0000
Sensitivity (%)	99.51	0.0049	0.7895
Specificity (%)	78.94	0.9782	1.0000
Positive case missing	2	1.0000	1.0000
Negative case missing	4		

B) CHAID

Summary of statistics		Points for plotting	
		FP rate	**TP rate**
Number of cases	431		
Number of correct	427	0.0000	0.0000
Accuracy	99.01	0.0000	0.0000
Sensitivity	99.76	0.0024	0.8421
Specificity	84.21	0.051	1.0000
Positive case missing	1	1.0000	1.0000
Negative case missing	3		

Figure 5. ROC curve information for C5.0 and CHAID models

Discussion

The present research aimed to improve diagnostic accuracy and personalized medicine by identifying differentially expressed genes (DEGs) that can serve as biomarkers for BC. Gene expression data were obtained from The Cancer Genome Atlas (TCGA) database and evaluated using supervised machine learning (ML) algorithms. The results indicated that both the C5.0 and CHAID models performed well in diagnosing BC, with the C5.0 model showing slightly better performance in the training data and the CHAID model performing better in the testing and validation phases. The CHAID model achieved 100% sensitivity and specificity in the validation data, with no false results. The study identified four important RNAs (ENSG00000171368, ENSG00000126218, ENSG00000123560, and ENSG00000228877) that are crucial for BC diagnosis.

Tubulin polymerization promoting protein (TPPP), ENSG00000171368, also known as p25, is one of the identified important classifiers in the diagnosis of BC (20). In terms of gene ontology, TPPP is enriched in the nucleus, cytoskeleton, microtubules, and mitotic spindles, and it exhibits metal ion binding, protein dimerization, and GTPase activities, as well as microtubule and tubulin formation. It encompasses various biological processes such as microtubule organization, cell proliferation, and regulation of protein-containing complex assembly. According to previous studies, TPPP is important in several conditions, including colorectal cancer development, oral squamous cell carcinoma, neurodegenerative diseases, oligodendrocyte differentiation, microtubule dysfunction in cystic fibrosis, and multiple sclerosis (21,22). Its importance in colorectal cancer progression is related to promoting cell proliferation, migration, and invasion. However, we could not find any previous study focusing on the expression changes of TPPP in bladder cancer.

Coagulation factor X, ENSG00000126218, encoded by the F10 gene, is another important classifier identified for BC diagnosis. In terms of gene ontology, coagulation factor X is involved in biological processes associated with the coagulation system, including regulation of blood coagulation and hemostasis. It is located in the extracellular space and blood microparticles and is functionally important in serine-type endopeptidase activity, calcium ion binding, and vitamin K binding. The gene encodes a serine protease that plays a crucial role in the coagulation cascade and hemostasis. Upon activation, it converts prothrombin to thrombin, which helps form fibrin clots to stop bleeding. We did not find any previous research indicating a direct association between expression levels of F10 and bladder cancer. A probable link between coagulation factors and cancer is the ability of cancer cells to induce a procoagulant state, resulting in activation of the coagulation cascade, which may support tumor spread and metastasis. A study on BC patients reported elevated fibrinogen levels compared to healthy individuals; suggesting that evaluating coagulation factors may potentially serve as screening or prognostic markers in cancer research (23).

Proteolipid Protein 1 (PLP1), also known as GPM6C and SPG2, is primarily known for its role in the central nervous system, particularly in myelination and acute myeloid leukemia (24). It is encoded by the PLP1 gene on the X chromosome (ENSG00000123560) and is one of the important classifiers identified in this study. In the context of bladder cancer, PLP1 is not typically investigated, however, a possible association with BC has been reported through bioinformatic analysis (25). The other important classifier identified is ENSG00000228877, also known as RP11-473E2. It refers to a long intergenic non-coding RNA whose functional role has not yet been fully identified.

Although the diagnostic models developed in this research demonstrated excellent performance metrics, it is important to acknowledge that these findings are exclusively based on in silico analyses and computational modeling using publicly available transcriptomic data. We did not validate these results through wet-lab or clinical experiments to confirm the biological relevance of the identified biomarkers in bladder cancer patients. However, this limitation does not undermine the robustness of our computational findings, rather, it highlights the need for future studies using experimental techniques. Further investigation will enhance our understanding across different pathological and physiological contexts.

Conclusion

In conclusion, the potential application of ML algorithms in identifying DEGs for BC diagnosis was investigated in this manuscript, which may contribute to better management and treatment outcomes for patients. Further investigations through laboratory validation of the results will provide deeper insights into the diagnosis of bladder cancer.

Acknowledgement

Not applicable.

Funding sources

There was no funding.

Ethical statement

This research is based on secondary analysis of existing, publicly available datasets. All data used in this study were obtained from databases that provide fully anonymized and aggregated information, with all personally identifiable details removed prior to public deposition. Therefore, this study did not involve direct interaction with human subjects and did not require review or approval by an Institutional Review Board (IRB) or Ethics Committee.

Conflicts of interest

The authors declare that they have no conflict of interest.

Author contributions

In the current research, each author made significant contributions to the study. M.A., as the corresponding author and project supervisor, contributed to the entire study process, including conceptualization, data collection, processing, interpretation, and manuscript revision. N.N., as the first author, contributed to conceptualization, study design, bioinformatic and machine learning workflow methodology, data interpretation, and manuscript drafting. E.M. contributed primarily to the clinical aspects of this project.

Data availability statement

This study is a secondary analysis of existing, de-identified, publicly available data. No new primary data were generated.

References

1. Koul B, Koul B. Types of Cancer. In: Herbs for Cancer Treatment. Singapore: Springer; 2019.p.53-150 [View at Publisher] [DOI] [Google Scholar]
2. Jubber I, Ong S, Bukavina L, Black PC, Compérat E, Kamat AM, et al. Epidemiology of bladder cancer in 2023: a systematic review of risk factors. Eur Urol. 2023;84(2):176-90. [View at Publisher] [DOI] [PMID] [Google Scholar]
3. Crabb SJ, Douglas J. The latest treatment options for bladder cancer. Br Med Bull. 2018;128(1):85-95. [View at Publisher] [DOI] [PMID] [Google Scholar]
4. Lopez-Beltran A, Cookson MS, Guercio BJ, Cheng L. Advances in diagnosis and treatment of bladder cancer. BMJ. 2024:384:e076743. [View at Publisher] [DOI] [PMID] [Google Scholar]
5. DeGEORGE KC, Holt HR, Hodges SC. Bladder cancer: diagnosis and treatment. Am Fam Physician. 2017;96(8):507-14. [View at Publisher] [PMID] [Google Scholar]
6. Ivanisevic T, Sewduth RN. Multi-omics integration for the design of novel therapies and the identification of novel biomarkers. Proteomes. 2023;11(4):34. [View at Publisher] [DOI] [PMID] [Google Scholar]
7. Kirk S, Lee Y, Lucchesi F, Aredes N, Gruszauskas N, Catto J, et al. The Cancer Genome Atlas Urothelial Bladder Carcinoma Collection (TCGA-BLCA)(Version 8)[Data set]. Cancer Imaging Arch. 2016;K9(10). [View at Publisher] [DOI]
8. Cortés-Ciriano I, Gulhan DC, Lee JJ-K, Melloni GE, Park PJ. Computational analysis of cancer genome sequencing data. Nat Rev Genet. 2022;23(5):298-314. [View at Publisher] [DOI] [PMID] [Google Scholar]

9. Shen J, Shi J, Luo J, Zhai H, Liu X, Wu Z, et al. Deep learning approach for cancer subtype classification using high-dimensional gene expression data. BMC bioinformatics. 2022;23(1):430. [View at Publisher] [DOI] [PMID] [Google Scholar]
10. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell. 2017;171(3):540-56.e25. [View at Publisher] [PMID] [Google Scholar]
11. Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. BMJ Open. 2021;11(12):e053674. [View at Publisher] [DOI] [PMID] [Google Scholar]
12. Al-Tashi Q, Saad MB, Muneer A, Qureshi R, Mirjalili S, Sheshadri A, et al. Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. Int J Mol Sci. 2023;24(9):7781. [View at Publisher] [DOI] [PMID] [Google Scholar]
13. Mazlan AU, binti Sahabudin NA, Remli MA, Ismail NSN, Mohamad MS, Abd Warif NB, editors. Supervised and unsupervised machine learning for cancer classification: recent development. 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS); 2021: IEEE. [View at Publisher] [DOI]
14. Li J, Wang Z, Wang T. Machine-learning prediction of a novel diagnostic model using mitochondria-related genes for patients with bladder cancer. Sci Rep. 2024;14(1):9282. [View at Publisher] [DOI] [PMID] [Google Scholar]
15. Liosis KC. Utilizing Machine Learning Methods for Genomic Biomarker Discovery in Prostate and Bladder Cancer. 2021. [View at Publisher] [Google Scholar]
16. Liosis KC, Marouf AA, Rokne JG, Ghosh S, Bismar TA, Alhajj R. Genomic Biomarker Discovery in Disease Progression and Therapy Response in Bladder Cancer Utilizing Machine Learning. Cancers. 2023;15(19):4801. [View at Publisher] [DOI] [PMID] [Google Scholar]
17. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g: Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51(W1):W207-W12. [View at Publisher] [DOI] [PMID] [Google Scholar]
18. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36(8):2628-9. [View at Publisher] [DOI] [PMID] [Google Scholar]
19. Berrar D. Performance measures for binary classification. In: Encyclopedia of Bioinformatics and Computational Biology. Vol. 1. Amsterdam: Elsevier; 2019.p.546-60. [View at Publisher] [DOI] [Google Scholar]
20. Chen Q, Yang C, Chen L, Zhang J-J, Ge W-L, Yuan H, et al. YY1 targets tubulin polymerisation-promoting protein to inhibit migration, invasion and angiogenesis in pancreatic cancer via p38/MAPK and PI3K/AKT pathways. Br J Cancer. 2019;121(11):912-21. [View at Publisher] [DOI] [PMID] [Google Scholar]
21. Li Y, Xu Y, Ye K, Wu N, Li J, Liu N, et al. Knockdown of tubulin polymerization promoting protein family member 3 suppresses proliferation and induces apoptosis in non-small-cell lung cancer. J Cancer. 2016;7(10):1189-96. [View at Publisher] [DOI] [PMID] [Google Scholar]
22. Xiao T, Lin F, Zhou J, Tang Z. The expression and role of tubulin polymerization-promoting protein 3 in oral squamous cell carcinoma. Arch Oral Biol. 2022;143:105519. [View at Publisher] [DOI] [PMID] [Google Scholar]
23. Alevizopoulos A, Tyritzis S, Leotsakos I, Anastasopoulou I, Pournaras C, Kotsis P, et al. Role of coagulation factors in urological malignancy: a prospective, controlled study on prostate, renal and bladder cancer. Int J Urol. 2017;24(2):130-6. [View at Publisher] [DOI] [PMID] [Google Scholar]
24. Xie F, Qu J, Lin D, Feng K, Tan M, Liao H, et al. Reduced Proteolipid Protein 2 promotes endoplasmic reticulum stress-related apoptosis and increases drug sensitivity in acute myeloid leukemia. Mol Biol Rep. 2023;51(1):10. [View at Publisher] [DOI] [PMID] [Google Scholar]
25. Wang H, Liu J, Lou Y, Liu Y, Chen J, Liao X, et al. Identification and preliminary analysis of hub genes associated with bladder cancer progression by comprehensive bioinformatics analysis. Sci Rep. 2024;14(1):2782. [View at Publisher] [DOI] [PMID] [Google Scholar]

Cite this article as:

Nematy N, Moudi E, Arabfard M. Effective diagnosis of bladder cancer: Leveraging bioinformatics and machine learning techniques. *Jorjani Biomedicine Journal.* 2025;13(X):X. http://dx.doi.org/10.29252/jorjanibiomedj.13.X.X