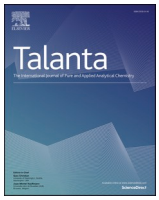
Talanta 297 (2026) 128749

Contents lists available at ScienceDirect

Talanta

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



Non-invasive bladder cancer detection: Identification of a urinary volatile biomarker panel using GC-MS metabolomics and machine learning

Â. Carapito [a,b,**], V.S. Fernandes Ferreira [c], A.C. Silva Ferreira [c,d,e], A. Teixeira-Marques [f,g], R. Henrique [f,g,h], C. Jerónimo [f,g], A.C.A. Roque [i,j], F. Carvalho [a,b], J. Pinto [a,b,1], P. Guedes de Pinho [a,b,1,*]

[a] Associate Laboratory i4HB - Institute for Health and Bioeconomy, University of Porto, Porto, Portugal
[b] UCIBIO – Applied Molecular Biosciences Unit, Laboratory of Toxicology, Faculty of Pharmacy, University of Porto, Porto, Portugal
[c] Cork Supply Portugal, S.A., Sao Paio de Oleiros, Portugal
[d] Centro de Biotecnologia e Química Fina (CBQF), Laboratório Associado, Escola Superior de Biotecnologia, Universidade Católica Portuguesa, Porto, Portugal
[e] Institute for Wine Biotechnology (IWBT), Department of Viticulture and Oenology (DVO), University of Stellenbosch, South Africa
[f] Cancer Biology and Epigenetics Group, Research Center (CI-IPOP), Portuguese Oncology Institute of Porto (IPO Porto) / Porto Comprehensive Cancer Center Raquel Seruca (Porto.CCC), Porto, Portugal
[g] Department of Pathology and Molecular Immunology, ICBAS-School of Medicine and Biomedical Sciences, University of Porto, Porto, Portugal
[h] Department of Pathology, Portuguese Oncology Institute of Porto (IPO Porto) / Porto Comprehensive Cancer Center Raquel Seruca (Porto.CCC), Porto, Portugal
[i] Associate Laboratory i4HB – Institute for Health and Bioeconomy, NOVA School of Science and Technology, NOVA University of Lisbon, Caparica, Portugal
[j] UCIBIO – Applied Molecular Biosciences Unit, Department of Chemistry, NOVA School of Science and Technology, NOVA University of Lisbon, Caparica, Portugal

ARTICLE INFO

Handling Editor: Qun Fang

Keywords:
Bladder cancer
Urinary biomarkers
Volatile organic compounds
Metabolomics
Machine learning
Gas chromatography-mass spectrometry
Early detection

ABSTRACT

Early detection of bladder cancer (BC) remains a major clinical challenge due to the limitations of current diagnostic methods, which are often invasive, expensive, or insufficiently sensitive, particularly for early-stage disease. Metabolomics approaches, when integrated with machine learning (ML) techniques, offer a powerful platform for identifying novel, non-invasive biomarkers. In this study, urinary volatile organic compounds (VOCs) were analysed from 87 BC patients and 90 age- and sex-matched cancer-free controls using headspace solid-phase microextraction coupled with gas chromatography–mass spectrometry (HS-SPME/GC-MS). Of the 90 VOCs identified, 27 were selected and used to train five ML algorithms—random forest (RF), support vector machine (SVM), partial least squares-discriminant analysis (PLS-DA), extreme gradient boosting (XGBoost), and k-nearest neighbors (k-NN). Model performance was evaluated using cross-validation and an independent validation set, with metrics including area under the curve (AUC), sensitivity, specificity, and accuracy. RF achieved the highest performance using all 27 features (AUC = 0.913; sensitivity, specificity, and accuracy = 85 %). After feature selection, an eight-VOC panel improved performance on the validation set (AUC = 0.872; sensitivity = 89 %, specificity = 92 %, accuracy = 91 %). The panel included ketones, aldehydes, a short fatty alcohol, and a phenol compound—seven elevated in BC, and one (acetone) decreased. This panel outperformed FDA-approved urinary assays and closely matched the specificity of urine cytology. These findings underscore the promise of VOC-based urinary biomarkers, in combination with ML, for the non-invasive detection of BC. Further large-scale validation studies are essential to confirm diagnostic utility and enable clinical translation.

* Corresponding author. Associate Laboratory i4HB and UCIBIOLaboratory of Toxicology, Department of Biological SciencesFaculty of Pharmacy, University of PortoRua Jorge Viterbo Ferreira, 2284050-313, Porto, Portugal.
** Corresponding author. Associate Laboratory i4HB and UCIBIOLaboratory of Toxicology, Department of Biological SciencesFaculty of Pharmacy, University of PortoRua Jorge Viterbo Ferreira, 2284050-313, Porto, Portugal.
E-mail addresses: up201804198@up.pt (Â. Carapito), pguedes@ff.up.pt (P. Guedes de Pinho).
[1] Contributed equally as senior authors.

https://doi.org/10.1016/j.talanta.2025.128749
Received 1 July 2025; Received in revised form 18 August 2025; Accepted 24 August 2025
Available online 26 August 2025

1. Introduction

Bladder cancer (BC) represents a significant global health concern, contributing to both incidence and mortality worldwide [1]. Current diagnostic methods for BC involve cystoscopy, urine cytology, and imaging techniques [2]. Cystoscopy remains the clinical gold standard, with a meta-analysis [3] reporting a sensitivity of around 75 % and a specificity of 74 %. However, cystoscopy is an invasive procedure and may skip certain lesions, particularly flat or small ones, such as carcinoma in situ [4]. Urine cytology, a non-invasive test that examines exfoliated cells in urine, is commonly used alongside cystoscopy. While it demonstrates high specificity (~95 %) [5], its overall sensitivity is low (~37 %) [5] and may reach only 24 % [6] for low-grade tumours, with accuracy depending heavily on specimen quality and the cytopathologist's experience [2]. Cross-sectional imaging (computed tomography, CT; and magnetic resonance imaging, MRI) is indispensable for staging muscle-invasive disease but is of limited utility in detecting superficial lesions and may yield false positives in benign conditions such as benign prostatic hypertrophy or bladder inflammation [7]. The high recurrence nature of BC requires active, ongoing surveillance strategies that are tailored to individual patient risk profiles [8,9]. For instance, patients classified as low-risk typically undergo cystoscopic evaluation at three- and twelve-months post-diagnosis. Surveillance is often reduced to annual monitoring thereafter if no recurrence is detected. High-risk patients, in contrast, undergo both cystoscopy and urinary cytology every three months for the first two years, every six months through year 5, and then shift to yearly exams [8,9]. Though effective at detecting recurrence, these intensive protocols expose patients to repeated invasive procedures, leading to discomfort and risks of complications such as bleeding, infection, and pain [8,10]. Additionally, the cumulative costs associated with frequent surveillance and treatment impose a substantial financial burden on healthcare systems [11].

To supplement these current methods, the U.S. Food and Drug Administration (FDA) has approved six urinary assays for BC detection and surveillance [12,13]. Among these, UroVysion (a fluorescence in situ hybridization assay detecting chromosomal abnormalities) and two assays for nuclear matrix protein 22 (qualitative "BladderChek" and quantitative ELISA) are intended for initial diagnosis [2,14,15]. While these tests offer the advantage of greater convenience and non-invasiveness, they are also limited by high false-positive rates, reduced specificity in patients with benign urological conditions, and poor sensitivity for low-grade tumours [12,13]. Consequently, despite these advances, there remains an unmet need for more accurate, non-invasive diagnostics that can reduce reliance on invasive cystoscopy and detect recurrence at an early, actionable stage.

Considering these challenges, exploring alternative tools, such as metabolomics-based approaches, is crucial for improving the effectiveness of BC diagnosis and ultimately enhancing patient outcomes. Metabolomics has demonstrated enormous potential to discover disease biomarkers and revolutionize our understanding of pathological processes. Its comprehensive approach involves analysing small (low molecular weight) metabolites in biological matrices (e.g., blood, urine, exhaled breath, faeces) [16,17]. In particular, gas chromatography–mass spectrometry (GC-MS) has emerged as a key analytical technique for profiling volatile organic compounds (VOCs) in biological samples. GC-MS enables the detection of subtle changes in metabolic profiles that are often associated with cancer, providing a readily accessible means for early disease detection [18]. VOCs are metabolic byproducts that reflect the biochemical changes occurring within the body [19]. Their distinctive odours vary based on molecular structure and concentration, often influenced by different physiological conditions. This variation may create unique scent signatures that serve as diagnostic biomarkers for certain diseases [20]. Trained animals, particularly dogs, possess remarkable olfactory abilities and have been shown to detect these changes in VOC profiles [21]. For example, dogs have identified "smell fingerprints" in biological samples associated with cancer [22], diabetes [23], and Parkinson's disease [24]. A notable case [25] in 1989 involved a dog detecting its owner's melanoma, further highlighting the potential of VOCs as biomarkers. However, the data generated by analytical techniques like GC-MS are complex and high-dimensional. To interpret and extract clinically relevant patterns, machine learning (ML) algorithms can be applied [26]. These algorithms effectively analyse multivariate data, identifying correlations and distinct VOC signatures that differentiate BC patients from cancer-free individuals [27]. This facilitates the development of robust, non-invasive diagnostic models. Both unsupervised and supervised ML approaches are useful [28]: unsupervised methods can reveal hidden patterns without prior knowledge of sample classification, while supervised methods such as partial least squares-discriminant analysis (PLS-DA), support vector machines (SVM), and random forests (RF) are typically preferred for their accuracy in classifying labelled samples and predicting outcomes [29].

A growing number of studies [30–36] have investigated urinary VOCs as potential biomarkers for BC detection using metabolomics. Most rely on headspace solid-phase microextraction coupled to GC-MS (HS-SPME/GC-MS) [30–32,34], though instrumentation varies (e.g., HS-SPME/GC × GC-TOF-MS [35] or HS-SPME/GC-TOF-MS with portable GC-ion mobility spectrometry [36]). These studies also vary in cohort size and the number and identity of candidate biomarkers reported. For instance, Pinto et al. [30] identified 17 altered volatiles with a combined sensitivity of 70 %, specificity of 89 %, and accuracy of 80 %. Lett et al. [31] reported a VOC panel of eight compounds yielding 71 % sensitivity and 72 % specificity. Ligor et al. [35] described 12 compounds exclusive to BC patients, while Cauchi et al. [34] achieved 90 % sensitivity and 88 % specificity using 16 candidate VOCs. Other studies, including those by Jobu et al. [32] and Tyagi et al. [36], identified smaller groups of five and 13 VOCs, respectively, with performance metrics ranging from 27 % sensitivity to 94 % specificity [36]. These findings demonstrate the promise of urinary VOCs as BC biomarkers but also underscore the variability in methods, cohort sizes, and statistical strategies.

Our study was designed to build upon and extend previous research in this area. In contrast to previous studies, we expanded the cohort size to include 177 urine samples, thereby enhancing the statistical power of our findings. Furthermore, while previous studies [30–36] primarily used chemometric methods, exploratory analyses, and some ML, we implemented more complementary ML classifiers, including RF, SVM, k-nearest neighbors (k-NN), and extreme gradient boosting (XGBoost). This comprehensive comparison allowed us to identify which algorithms most reliably distinguish BC cases from controls and to develop a consensus panel of VOCs that maintained high predictive performance across models. We employed HS-SPME/GC-MS to generate a high-dimensional profile of urinary VOCs and volatile carbonyl compounds (VCCs) - a subclass of VOCs including aldehydes and ketones. These complex data were then subjected to an ML workflow to identify a concise panel of urinary biomarkers that consistently deliver high predictive performance across models. By integrating robust analytical chemistry with a comprehensive ML framework, this work aims to yield a non-invasive biomarker panel capable of early BC detection.

2. Materials and methods

2.1. Chemicals

4-Methyl-2-pentanone (≥99 %), nonanal (95 %), 2-ethyl-1-hexanol (≥99.6 %), 2,5-dimethylbenzaldehyde (99 %), hexanal (98 %), *p*-cresol (≥99 %), *m*-cresol (≥99 %), *o*-cresol (≥99 %), and O-(2,3,4,5,6-pentafluorobenzyl)-hydroxylamine hydrochloride (PFBHA) (≥99 %) were purchased from Merck (Darmstadt, Germany). 4-Heptanone (≥97 %), cyclohexanone (≥99 %), benzaldehyde (≥99.5 %), methylglyoxal (40 %, aqueous solution), and formaldehyde (37 %, aqueous solution) were purchased from Sigma-Aldrich (Madrid, Spain). Glyoxal (40 %,

aqueous solution), and acetaldehyde (≥99.5 %) were purchased from Fluka (Madrid, Spain). Acetone (≥99.7 %) was purchased from Labor-Spirit (Loures, Portugal). Sodium chloride was purchased from VWR (Leuven, Belgium).

2.2. Subject recruitment and urine collection

Volunteers for this study were recruited at the Portuguese Institute of Oncology of Porto (IPO Porto) and the Faculty of Pharmacy of the University of Porto. This study received approval from the Ethics Committee of IPO Porto (CES 82/022) and the Ethics Committee of the Faculty of Pharmacy of the University of Porto (CEFFUP). Urine samples from 177 individuals were collected in the morning (without fasting), including 87 patients diagnosed with BC and 90 cancer-free individuals (control group). All individuals signed an informed consent before entering the study. The control group consisted of individuals without a diagnosis of cancer but with age-related comorbidities (e.g., hypertension, diabetes, hypercholesterolemia). Data on the smoking habits of both groups were collected. Urine samples were matched between the BC and control groups for age and sex. BC and control groups were randomly divided into two different sets: a training set consisting of 70 % of the samples ($n = 61$ BC, and $n = 63$ controls), and a validation set consisting of 30 % of the samples ($n = 26$ BC, and $n = 27$ controls). Demographic and clinical data of BC patients and controls are shown in Table 1.

A total of 156 urine samples were initially collected and processed according to an ultracentrifugation protocol [37]. Urine was collected in sterile 120 mL polypropylene containers and immediately transferred into 50 mL Falcon tubes for an initial centrifugation at 3700×*g* for 20 min at 4 °C. The resulting pellets were discarded, and the supernatants were aliquoted into 2 mL Eppendorf tubes and kept at −80 °C. For ultracentrifugation processing, these supernatants were further centrifuged at 3000×*g* for 15 min at 4 °C, filtered through 0.22 μm PVDF membrane filters (Millipore, MA, USA), and ultracentrifuged at 200, 000×*g* for 48 min at 10 °C. The final supernatants were aliquoted into 2 mL Eppendorf tubes and stored at −80 °C until analysis. To expand the cohort size, an additional 21 urine samples ($n = 4$ controls and $n = 17$ BC), which had been processed using a standard centrifugation protocol (3700×*g*, 20 min, 4 °C), were also included.

Table 1
Demographic and clinical data of cancer-free controls and BC patients.

Group		Cancer-free controls		BC patients	
		Training set	Validation set	Training set	Validation set
Number of samples (F/M)		63 (17/46)	27 (7/20)	61 (18/43)	26 (7/19)
Age range (years)		45–90	45–90	42–87	43–86
Mean age ± SD (years)		62.5 ± 11.6	63.8 ± 12.4	66.7 ± 9.9	66.4 ± 11.9
Smoking habits					
Current smoker		7	2	32	13
Former smoker		9	7	8	2
Never smoker		21	8	5	4
Unknown		26	10	16	7
Stage and grade					
NMIBC	Ta	–	–	25	10
	T1	–	–	14	10
MIBC	T2	–	–	10	2
	T3	–	–	8	2
	T4	–	–	4	2
LG		–	–	16	7
HG		–	–	44	14
Unknown grade		–	–	1	5

BC bladder cancer, *F* female, *M* male, *SD* standard deviation, *NMIBC* non-muscle invasive bladder cancer, *MIBC* muscle-invasive bladder cancer, *LG* low-grade, *HG* high-grade.

2.3. Sample preparation

Urine samples were prepared according to two protocols previously optimized [30,38]. Urine samples were thawed at room temperature. One of the protocols involved the analysis of VOCs, by adding 0.54 g of NaCl to 2 mL of urine in a 10 mL glass vial. A second protocol focusing on the analysis of VCCs was also carried out, requiring prior derivatization of the carbonyl groups. In this procedure, 7.5 μL of O-(2,3,4,5,6-pentafluorobenzyl) hydroxylamine hydrochloride (PFBHA) (40 g/L solution) was added to 250 μL of urine in a 10 mL glass vial. VOCs and VCCs were extracted from urine samples by HS-SPME using conditions previously established in our laboratory [30,38,39]. For VOCs, each urine sample was first incubated for 11 min at 44 °C with continuous stirring (250 rpm) in a Combi-PAL autosampler (Varian Pal Autosampler, Zwingen, Switzerland). Then, the compounds were extracted from the headspace for 30 min (44 °C, 250 rpm) using a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fibre (Supelco Inc., Bellefonte, PA, USA), and thermally desorbed into the GC system for 4 min at 250 °C. In the VCCs protocol, each sample was incubated for 6 min at 62 °C with continuous stirring (250 rpm) in a Bruker CTC PAL-xt autosampler (Bruker Daltonics, Bremen, Germany). Derivatized molecules were then extracted from the headspace for 51 min (62 °C, 250 rpm) using a polydimethylsiloxane/divinylbenzene (PDMS/DVB) fibre (Supelco Inc., Bellefonte, PA, USA). The thermal desorption of VCCs into the GC injector was carried out for 5 min at 250 °C. Samples were injected randomly for both methods, and quality control (QC) samples were consistently injected after every ten samples.

2.4. GC-MS analysis and metabolite identification

A SCION single quadrupole (SQ) mass spectrometer (Bruker Daltonics, Bremen, Germany) coupled to a 436-GC model was used for VOC analysis. The GC system was equipped with a fused silica capillary column (Rxi-5Sil MS, 30 m × 0.25 mm internal diameter × 0.25 μm; Restek Corporation, U.S., Bellefonte, PA, USA) and high purity helium C-60 (Gasin, Porto, Portugal) as a carrier gas, operating at a constant flow rate of 1.0 mL/min. The oven temperature was programmed at 40 °C for 1 min, increased at 5.0 °C/min to 250 °C where it was held for 5 min, and then increased at 5.0 °C/min to 300 °C. The transfer line, ion source, and manifold temperatures were maintained at 250, 260, and 41 °C, respectively. The SQ-MS was operated in electron impact (EI) mode (70 eV). Data acquisition was performed in full scan mode from *m/z* 40 to 400 with a scan time of 500 ms. A 436-GC model coupled to an EVOQ triple quadrupole (TQ) mass spectrometer (Bruker Daltonics, Bremen, Germany) was used for VCC analysis. The column and the carrier gas were the same as described in the VOC analysis. The oven temperature was programmed at 40 °C for 1 min, increased at 5.0 °C/min to 250 °C where it was held for 5 min, and then increased at 20.0 °C/min to 300 °C. The transfer line, ion source, and manifold temperatures were maintained at 260, 270, and 41 °C, respectively. The TQ-MS was operated in EI mode (70 eV), and data acquisition was performed in full scan mode from *m/z* 35 to 600 with a scan time of 250 ms. The software used in both instruments was a Bruker Daltonics MS workstation (version 8.2.1, Bruker Daltonics, Bremen, Germany).

Metabolites were identified by comparing the mass spectra and retention indices (RI) of the chromatographic peaks in urine samples with those obtained from the National Institute of Standards and Technology (NIST) standard reference database (version 14, Gaithersburg, MD, USA). A tolerance of ±20 was defined to accept the difference between the experimental and theoretical RI from NIST, and a minimum reverse match of 750 was considered. Identification was further verified by analysis of commercial standard compounds under the same conditions. This procedure allowed for the establishment of confidence levels in metabolite identification, as recommended for metabolomic studies [40,41].

2.5. Statistical analysis

2.5.1. Data pre-processing

After converting the GC-MS raw data files from VOCs and VCCs to netCDF format, they were imported into PARADISe (v.6.0.1) software [42]. Data pre-processing comprised all steps performed to transform the raw chromatographic data into clean, structured datasets ready for statistical analysis. The first pre-processing step involved automated data alignment, essential for managing complex and large datasets, particularly when significant shifts occur during interval selection. Once alignment was complete, the next phase entailed manually selecting retention time intervals for each peak or co-eluted peak, allowing for precise individual modelling. This was succeeded by PARAFAC2 modelling and fitting, which deconvoluted the peaks and facilitated the putative identification of compounds by matching mass spectra of the PARAFAC2 components with the NIST mass spectral library. Ultimately, a comprehensive report was generated, exporting a final matrix that detailed the peak areas of the selected components. The exported PARADISe matrices included a report detailing the peak integration (peak areas) for all the exported peaks from each sample, along with information on the intervals and models, as well as the most suitable chemical identification based on the NIST match factor (the highest match factor). Another sheet also provides the *m/z* spectra for each exported peak [42].

During the selection step in PARADISe, artifact peaks originating from the fibre and the chromatographic column, such as siloxanes and cyclosiloxanes, were sistematically excluded. These were recognized based on mass spectra exhibiting characteristic fragment ions (*m/z* 73, 147, 207, 281), high match factors with siloxane-based compounds in the NIST library, and their consistent presence in blanks and QC samples at comparable or higher abundance than in urine samples. Finally, the data were normalized to the total signal area to ensure comparability across samples. Among the 90 compounds identified using both protocols (comprising 64 VOCs and 26 VCCs), ten were detected in both VOC and VCC chromatograms, resulting in duplicates. For each duplicate compound, the entry with the lower relative standard deviation (RSD, calculated from the QC samples as the standard deviation of the peak area divided by the mean peak area, and expressed as a percentage) and higher signal-to-noise ratio (S/N, calculated as the peak height divided by the baseline noise level) was retained. This refinement yielded a final set of 80 unique compounds for analysis. Of these, 46 compounds with an S/N below 10 were excluded, reducing the dataset to 34 compounds. Additionally, seven compounds known to originate from exogenous sources were removed, yielding two final matrices: one containing 14 VOCs and the other 13 VCCs. These normalized matrices were then auto-scaled before multivariate analysis.

2.5.2. Data processing

Data processing encompassed the analytical and statistical procedures performed on the pre-processed datasets to build predictive models and evaluate performance. First, the reproducibility of both analytical protocols was assessed by principal component analysis (PCA) in MetaboAnalyst 6.0 [43] using all samples ($n = 90$ BC, $n = 87$ cancer-free controls) and their respective QCs ($n = 15$ for VOCs; $n = 14$ for VCCs) (Fig S-1). Upon confirming reproducibility, the normalized VOC and VCC matrices were concatenated, resulting in a single dataset comprising 27 compounds.

The dataset was divided into training (70 %) and validation (30 %) sets. To ensure balanced representation of the target classes, we matched participants based on age and sex. To develop a robust and generalizable model capable of accurately identifying patterns and relationships in the data between BC and cancer-free controls, five different ML algorithms were implemented, including SVM, RF, PLS-DA, k-nearest neighbors (k-NN), and extreme gradient boosting (XGBoost). ML analyses were performed using Python software (https://www.python.org/), with libraries such as pandas [44], numpy [45], scikit-learn [46] and xgboost [47], with visualization supported by matplotlib [48] and seaborn [49]. All ML algorithms were implemented in Python. Hyperparameter optimization for each model was first performed on the training set using Grid Search [46] within a 3-fold cross-validation framework. This combined approach made the tuning of model parameters more reliable and reduced the risk of overfitting to a specific training or validation set, helping to create a model that performs well on new, unseen data. Next, the ML models were applied to the individual compounds that were shown to have the highest feature importance rank for discriminating between BC patients and cancer-free individuals. This ranking process allowed us to identify a biomarker panel with the highest discriminatory power by selecting the SelectKBest [50] ANOVA F-test (features with the highest importance scores). Following training set, the performance of the established panel was evaluated using an independent validation set, allowing for a comparative assessment of outcomes across multiple ML algorithms. Classification performance was assessed by calculating sensitivity as the number of true positives divided by the sum of true positives and false negatives, specificity as the number of true negatives divided by the sum of true negatives and false positives, and accuracy as the sum of true positives and true negatives divided by the total number of samples. Performance metrics also included the area under the curve (AUC) from ROC analysis, for both the training and validation sets to comprehensively assess and compare the predictive abilities of the models. Finally, univariate analysis was performed on the biomarker-panel metabolites using the unpaired Mann-Whitney test (statistical significance, p-value <0.05) in GraphPad Prism (version 8.0.2, San Diego, CA, USA), and boxplots were also generated to visually display the distribution of these compounds in the BC group versus the control group.

3. Results

This study evaluated the potential of urinary volatile profiles to distinguish between patients with BC ($n = 87$) and cancer-free controls ($n = 90$). The cohort consisted of 62 male and 25 female BC patients, reflecting the higher prevalence of BC among men. Most BC patients were current smokers ($n = 45$), in contrast to cancer-free individuals (only 9 were current smokers). Notably, most cancer-free individuals had never smoked ($n = 29$). Regarding BC stage, most patients were diagnosed with non-muscle invasive bladder cancer (NMIBC) ($n = 59$), while a smaller number were diagnosed with more advanced stages of muscle-invasive BC, specifically stages T3 ($n = 10$) and T4 ($n = 6$). As for grade, most BC patients received a diagnosis of high-grade BC ($n = 58$), compared to those diagnosed with low-grade BC ($n = 23$). Most of the compounds identified belonged to the classes of ketones ($n = 13$) and aldehydes ($n = 11$). The remaining compounds ($n = 3$) were a short fatty alcohol, phenol, and fatty acid. Further details regarding retention time (RT), the most abundant ions (*m/z*), experimental and theoretical retention indices (RI), R-match, CAS number, RSD based on repeated analysis of QC samples, and the identification level are provided in Supplementary Table S-1.

3.1. ML methods for BC detection

To verify the analytical reproducibility, a PCA was performed for VOCs and VCCs using the QC samples and all samples under study, as illustrated in Fig S-1. The clustering of the QC samples indicated a good reproducibility of the analytical methodology.

The dataset comprising the 27 identified volatile metabolites (all features) was used to train and evaluate several ML algorithms – namely RF, k-NN, SVM, PLS-DA, and XGBoost – to discriminate between BC and control samples. For each algorithm, performance metrics were derived using the entire dataset with all features. Feature importance rankings were generated for all the models, allowing the identification of the subset of compounds with higher importance for discrimination. Confusion matrices were generated for each model to evaluate their

diagnostic performance, providing insight into sensitivity, specificity, and overall classification accuracy. These metrics are summarized in Table 2, which details the comparative performance of the five algorithms.

For each algorithm, three sets of metrics are presented: performance using all 27 features on the training set, performance of the optimized model applied to the training set, and performance of the optimized model on the independent validation set. When using all features, the RF model yielded the best predictive performance, achieving a sensitivity, specificity, and accuracy of 85 %, with an AUC of 0.913 (Table 2). XGBoost achieved identical values for sensitivity, specificity and accuracy but reported a lower AUC. The SVM and PLS-DA models demonstrated comparable performance, with a sensitivity of 85 %, a specificity of 77 %, and an accuracy of 81 %, with AUCs of 0.811 and 0.830, respectively. In contrast, the k-NN algorithm exhibited the lowest performance among the ML algorithms, with a sensitivity of 78 %, a specificity of 65 %, an accuracy of 72 %, with an AUC of 0.716. After feature selection and hyperparameter optimization, each algorithm's best-performing model was identified and re-evaluated on the training set. All algorithms improved under this approach. XGBoost exhibited perfect performance on the training data, which may indicate a tendency toward overfitting. Validation set results revealed the practical performance of each model. Given the higher AUC and overall performance metrics, the RF algorithm remained the top-performing approach with a sensitivity of 89 %, specificity of 92 %, accuracy of 91 %, and an AUC of 0.872. SVM and PLS-DA yielded similar AUCs (0.830), with SVM showing slightly better balance between sensitivity and specificity. K-NN maintained low performance on validation, consistent with its results using all features. This RF model demonstrated the most robust performance, showing strong accuracy not only on the training data but also when applied to independent test data, indicating good generalizability. This model also allowed the identification of an eight-biomarker panel including three ketones (4-methyl-2-heptanone, 4-methyl-2-pentanone, and acetone), three aldehydes (nonanal, hexanal, and 2,5-dimethylbenzaldehyde), one short fatty alcohol (2-ethyl-1-hexanol), and one phenol compound (*p*-cresol). Seven of the eight selected biomarkers were present in higher levels in BC samples compared to cancer-free controls, while one biomarker was present in lower levels (Fig S-2). The variation values reported for each compound represent univariate measures of variance, reflecting the differences in metabolite levels between BC and control groups and the measurement uncertainty associated with each metabolite. These variations differ across compounds depending on how distinctly their relative abundance separate the two groups. By contrast, the feature importance scores were derived using the SelectKBest method with ANOVA F-test rankings within the RF model, indicating each compound's contribution to the overall classification performance when considered as part of the combined biomarker panel. Consequently, metabolites showing high univariate variation may not necessarily contribute strongly to group discrimination if their differences lack consistency or specificity, whereas others with moderate variation may achieve higher importance by exerting a synergistic effect within the multivariate model. The focus of this study was therefore to establish a biomarker panel that collectively yields the highest performance for early BC detection, rather than relying solely on individual metabolite variation. The ROC curve, confusion matrix, and feature importance using the SelectKBest method based on ANOVA F-test scores for the RF model are presented in Fig. 1, while corresponding results for the other algorithms are provided in Fig S-3. Table 3 presents the eight biomarker-panel metabolites estimated by the RF model, along with their respective feature importance scores. Additionally, Supplementary Table S-1 provides a comprehensive list of the 27 volatile metabolites that were identified and used in the ML analysis.

4. Discussion

In this study, we defined an eight-biomarker panel of urinary VOCs that demonstrated potential for BC detection. Among the candidate biomarkers identified in this study, seven compounds depicted higher levels in the urine of BC patients compared to cancer-free controls (4-methyl-2-heptanone, 4-methyl-2-pentanone, nonanal, 2-ethyl-1-hexanol, 2,5-dimethylbenzaldehyde, hexanal, and *p*-cresol). Conversely, one compound (acetone) was found to be present in lower levels. Among the eight compounds in the biomarker panel, six (4-methyl-2-heptanone [36], nonanal [31,36], hexanal [30,34], 2-ethyl-1-hexanol [31], 2, 5-dimethylbenzaldehyde [35], and p-cresol [30]) had been previously reported to vary significantly between BC and controls, whereas two compounds - acetone and 4-methyl-2-pentanone – were not previously associated with BC.

The eight-biomarker panel includes three aldehydes, namely nonanal, 2,5-dimethylbenzaldehyde, and hexanal. Nonanal has been associated with BC [31,36] as well as with other cancer types, such as renal cell carcinoma (RCC) [52] and head and neck squamous cell carcinoma [53]. In the context of BC, nonanal has been identified as a potential biomarker in the study by Tyagi et al. [36] (no tendency specified). However, another study reported decreased levels of nonanal in BC [31], in contrast to our findings, where higher levels of nonanal were observed in urine of BC patients in comparison with controls. Elevated levels of nonanal have also been detected in the urine of RCC and head and neck squamous cell carcinoma patients compared to control groups. In the study by Wang et al. [52], the increased presence of nonanal and other aldehydes in RCC patients suggested that oxidative stress may contribute to the production of these compounds. Likewise, Opitz et al. [53] demonstrated that aldehydes, such as nonanal, are byproducts from lipid peroxidation of unsaturated fatty acids, a process closely linked to oxidative stress. This stress may result from an imbalance between free oxygen radicals and antioxidants, leading to the overproduction of reactive oxygen species (ROS), which play a critical role in various pathological processes, including carcinogenesis.

In our study, we observed that 2,5-dimethylbenzaldehyde was present in higher levels in the urine of BC patients compared with controls, which aligns with the findings of Ligor et al. [35] Similarly, Lima et al. [38] reported the presence of higher levels of this compound in the urine of prostate cancer (PCa) patients in comparison with controls.

Hexanal has been widely associated with various cancer types, such as pancreatic ductal adenocarcinoma [54], BC [30,34], head and neck squamous cell carcinoma [53], lung cancer [55–58], and colorectal,

Table 2
Performance of ML models used for BC detection based on the training set using all features, the best model on the training set, and the best model on the validation set.

Algorithm	All 27 features (Training set)				Best model (Training set)				Best model (Validation set)			
	AUC	Sens.	Spec.	Accu.	AUC	Sens.	Spec.	Accu.	AUC	Sens.	Spec.	Accu.
RF*	0.913	85	85	85	0.960	95	97	96	0.872 (8 features)	89	92	91
XGBoost	0.849	85	85	85	1	1	1	100	0.849 (27 features)	85	85	85
SVM*	0.811	85	77	81	0.900	89	79	84	0.830 (5 features)	85	81	83
PLS-DA*	0.830	85	77	81	0.914	94	75	85	0.830 (27 or 12 features)	85	77	81
k-NN	0.716	78	65	72	0.968	97	87	92	0.716 (5 features)	78	65	72

Accu. Accuracy, *AUC* area under the curve, *RF* random forest, *k-NN* k-nearest neighbors, *PLS-DA* partial least squares-discriminant analysis, *Sens.* Sensitivity, *Spec.* specificity, *SVM* support vector machine, *XGBoost* extreme gradient boosting. *with optimization and scaling.

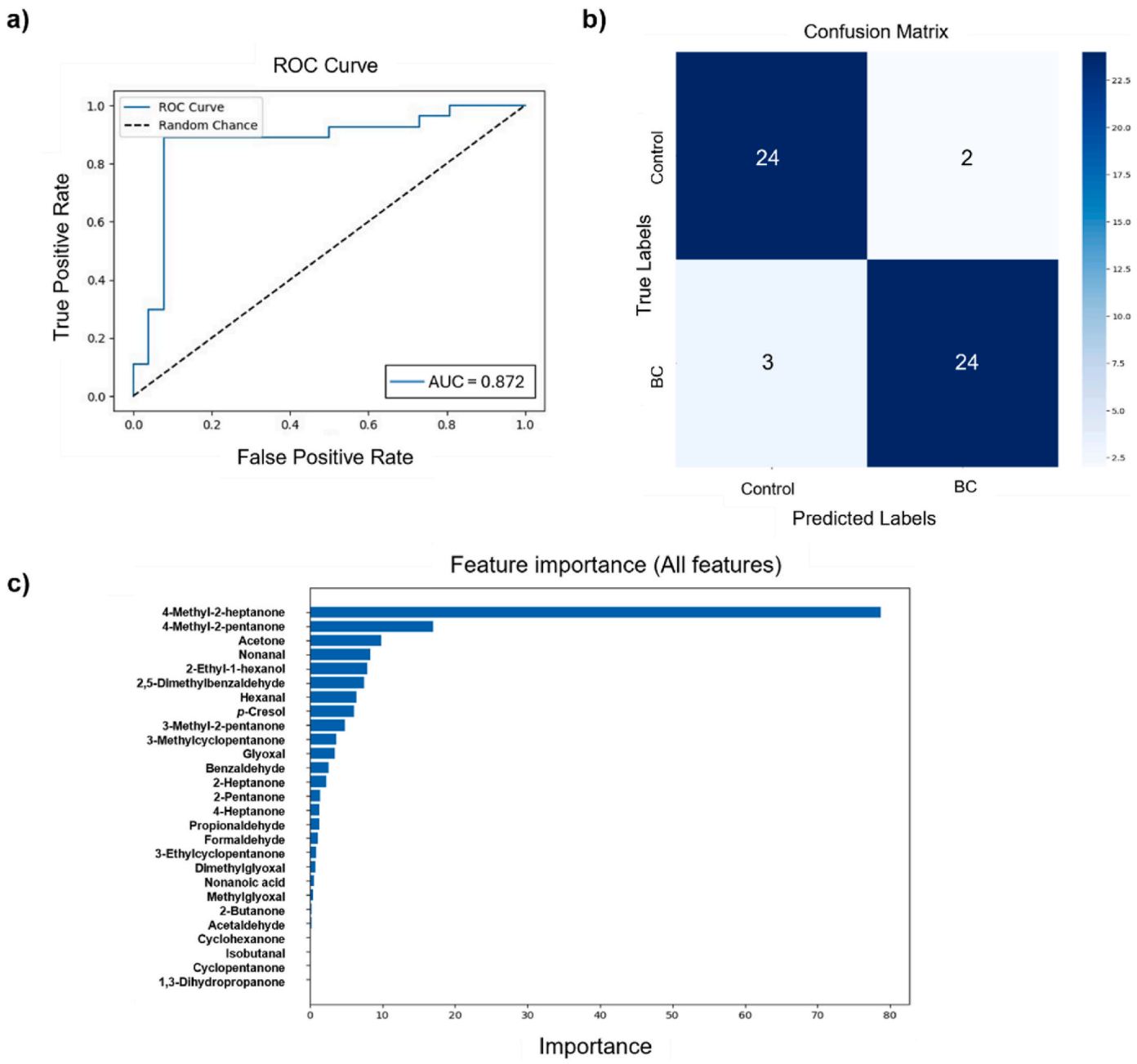


Fig. 1. a) ROC curve, **b)** confusion matrix, and **c)** feature importance scores of urinary VOCs estimated using an RF model. Input features were preselected using the SelectKBest method based on ANOVA F-test scores. The feature importance plot shows all 27 features ranked by importance, with the top eight metabolites considered for the biomarker panel.

leukaemia, and lymphoma [59]. Higher levels of hexanal in the urine of cancer patients were reported in six studies [34,53,55,57,58,60], while lower levels were reported in three studies [30,54,59]. In our study, hexanal was found at higher levels in the urine of BC patients compared with controls, a finding consistent with the results reported by Cauchi et al. [34]. However, this contrasts with another study [30] that reported lower levels of hexanal in BC. Several factors may account for these differences. One possibility is the variability in the activity of aldehyde dehydrogenase (ALDH), the enzyme responsible for converting hexanal to its carboxylic acid. An upregulation of ALDH in certain BC cases could lead to reduced hexanal levels [30], while decreased ALDH activity may result in higher levels of this compound [18]. Furthermore, differences in the extent of lipid peroxidation [30] – which reflects varying levels of oxidative stress and subsequent membrane degradation – could also impact the levels of hexanal produced [18].

The eight-biomarker panel also comprises three ketones, including 4-methyl-2-heptanone, 4-methyl-2-pentanone, and acetone. In our study, we found that 4-methyl-2-pentanone was present at higher levels in the urine of BC patients compared with controls. This compound was previously identified as a potential urinary biomarker of BC in the study of Tyagi et al. [36] (no tendency specified). Notably, it has not been recognized as a potential biomarker in the urine of patients with other cancer types. 4-Methyl-2-heptanone was identified in our study at higher levels in the urine of BC patients compared with controls. To the best of our knowledge, this is the first time that 4-methyl-2-heptanone is shown as a potential BC biomarker. Previously, its higher levels were only reported in the urine of patients with head and neck squamous cell carcinoma [53]. Regarding acetone, our study found lower levels of this

Table 3
List of the eight metabolites selected for the biomarker panel by the RF model, along with their respective feature importance scores.

Compound name[a]	Effect size ± ES_{SE}[b]	Variation ± uncertainty (%)	Feature importance[c]	Down- or upregulated
4-Methyl-2-heptanone (4-methylheptan-2-one) [d, g]	1.53 ± 0.34	188.9 ± 10.8	78.7	↑
4-Methyl-2-pentanone (4-methylpentan-2-one) [d, f]	0.76 ± 0.31	51.7 ± 9.1	16.9	↑
Acetone (propan-2-one) [d, f]	−0.53 ± 0.30	−26.3 ± 9.4	9.8	↓
Nonanal [d, f]	0.51 ± 0.30	166.4 ± 30.2	8.4	↑
2-Ethyl-1-hexanol (2-ethylhexan-1-ol) [d, f]	0.50 ± 0.30	81.8 ± 19.9	7.9	↑
2,5-Dimethylbenzaldehyde [e, f]	0.43 ± 0.30	60.1 ± 18.0	7.5	↑
Hexanal [d, f]	0.34 ± 0.30	28.4 ± 12.2	6.4	↑
p-Cresol (4-methylphenol) [d, f]	0.42 ± 0.30	385.9 ± 53.5	6.1	↑

[a] Common metabolite name (IUPAC name).
[b] Effect size ± ES_{SE} (effect size standard error) determined as described in the reference [51].
[c] Feature importance based on SelectKbest f-test.
[d] [e] Compounds detected through VOCs and VCCs analytical methods, respectively.
[f] Identified metabolites (confirmed using a chemical reference standard).
[g] Putatively annotated compounds (spectral MS similarity with the NIST database).

compound in the urine of BC patients compared with controls. This compound has not been previously associated with BC as a potential biomarker. Two studies found increased urinary levels of acetone in gastro-esophageal [61] and head and neck squamous cell carcinoma [53] patients, whereas another study reported decreased levels in the urine of lung cancer patients [55]. Higher levels of acetone may reflect the underlying cancer status or physiological adaptations, such as malnutrition, that shift metabolism towards ketogenic processes [61]. In the human body, acetone is primarily produced through the decarboxylation of acetoacetate, a ketone body generated during lipolysis and the catabolism of ketogenic amino acids. In cancer, metabolic reprogramming – such as the Warburg effect – leads to increased glycolysis and reduced reliance on the Krebs cycle, resulting in the accumulation of excess acetyl coenzyme A [18]. This surplus is diverted toward ketogenesis in the liver, enhancing the production of ketone bodies, including acetone. As a result, elevated levels may be observed in cancer patients, reflecting the metabolic shifts associated with tumour progression [18]. However, if ketogenesis is downregulated, due to sufficient glucose availability or impaired fatty acid oxidation, less acetone is generated. Also, acetone can be further metabolized by enzymes such as CYP2E1 into acetol, resulting in its lower levels [18].

The remaining two compounds in the eight-biomarker panel belong to the fatty alcohol and phenol classes, namely 2-ethyl-1-hexanol and *p*-cresol, respectively. In our study, 2-ethyl-1-hexanol was present at higher levels in the urine of BC patients compared with controls. Interestingly, elevated levels of 2-ethyl-1-hexanol have also been observed in patients with lung [62] and breast [63] cancers, however reduced levels have been reported not only in BC [31] but also in two studies involving PCa [64,65]. In the study by Liu et al. [65] decreased levels of this compound were found in PCa patients, leading the authors to suggest that lower levels of 2-ethyl-1-hexanol may impair PCa cell apoptosis, potentially facilitating the progression of the disease. In cases where increased levels of 2-ethyl-1-hexanol are observed, the findings suggest a metabolic shift characterized by enhanced conversion of hydrocarbons into fatty alcohols [18,66]. This process is mediated by ADH and cytochrome P450 enzymes, both of which play a crucial role in metabolizing aldehydes and ketones derived from hydrocarbons [18]. Therefore, the elevated levels of 2-ethyl-1-hexanol likely reflect a response to increased oxidative stress and altered enzymatic activity in cancer cells [67].

p-Cresol has been found in our study at higher levels in the urine of BC patients compared with controls. Similarly, this trend has been reported in several studies involving different types of cancer, such as BC [30], lung [55], breast [66,68], colon [68], colorectal, leukaemia, and lymphoma [59]. In fact, *p*-cresol has been found altered in the urine of cancer patients with the same tendency (increased) throughout studies. *p*-Cresol is primarily produced by the gut microbiota through the fermentation of aromatic amino acids, namely tyrosine and phenylalanine [69]. Once synthesized, it undergoes hepatic conjugation via sulphation and glucuronidation, facilitating its excretion primarily through urine [69]. This increase may reflect alterations in gut microbial composition associated with BC, potentially leading to enhanced production of *p*-cresol. Notably, *p*-cresol has been implicated in promoting cancer cell migration and invasion via activation of Ras and mTOR signalling pathways, suggesting that elevated urinary *p*-cresol levels may not only serve as a biomarker but also contribute to tumour progression [70].

The incorporation of VOCs into specific biochemical pathways is a highly challenging process due to the considerable diversity of VOCs, their multiple potential sources, and the inherent complexity of metabolic networks. These factors result in insufficient data for meaningful pathway enrichment analysis and make it difficult to establish clear associations. Additionally, individual metabolic variability, environmental influences, and intricate biological interactions further hinder this association [18]. Moreover, the participation of VOCs in metabolic pathways remains difficult to elucidate due to their limited exploration in current research [18].

Regarding the performance of our eight-biomarker panel, it was compared to traditional methods and three FDA-approved urinary biomarker tests for the detection of BC. Our panel demonstrated superior performance in this regard. Specifically, it achieved a sensitivity of 89 %, which is higher than that of all standard and FDA-approved tests (see Table 4). In terms of specificity, our panel reached 92 %, a value that was only exceeded by urine cytology, which reported a specificity of 93 %.

One limitation of our study is the small number of samples available for each disease stage and grade (Table 1), which prevented the use of ML algorithms. As a result, robust comparisons between these subgroups and the control group could not be performed. Another limitation was the imbalance in smoking status within the control group, with a small number of current smokers ($n = 9$) compared to individuals who never smoked ($n = 29$), which may have limited the ability to explore the influence of smoking on urinary VOC profiles. Nevertheless, our findings demonstrated that BC patients and cancer-free individuals exhibit

Table 4
Performance of standard diagnostic tests, FDA-approved tests, and the set of eight-biomarkers for BC detection.

Test	Sensitivity (%)	Specificity (%)
Cystoscopy	75 [3]	74 [3]
Urine cytology	37 [5]	95 [5]
UroVysion (FISH)	63 [2]	87 [2]
NMP22 (ELISA)	69 [2]	77 [2]
NMP22 BladderCheck (POC)	58 [2]	88 [2]
Eight-biomarker panel[a]	89	92

ELISA enzyme-linked immunosorbent assay, *FISH* fluorescence in situ hybridization, *NMP22* nuclear matrix protein 22, *POC* point-of-care.
[a] Performance achieved by the RF model on the validation set.

distinct urinary volatile profiles. By applying different ML algorithms, we identified a panel of eight compounds that, when combined, achieved high performance for BC detection with a sensitivity of 89 %, a specificity of 92 %, and an accuracy of 91 %. Importantly, establishing a panel of compounds, rather than relying on a single biomarker, is crucial because individual compounds may not show the ability to discriminate cancer patients from cancer-free individuals when assessed alone. In contrast, the combination of multiple compounds can enhance detection performance and provide a more robust BC detection signature. These results underscore that VOC profiling through metabolomics offers a promising, non-invasive approach for early BC detection by analysing specific compounds in urine. However, to translate this eight-biomarker into clinical practice, further work is needed to establish robust quantification and standardization. Key steps include: 1) accurate quantification of the volatile compounds in the panel to obtain detailed concentration data that distinguish BC from control samples, using one or more internal standards to control for variability in sample preparation, extraction, and instrument performance over time, providing consistent results regardless of the laboratory performing the analysis; 2) algorithm development using these quantitative data to classify samples based on the concentration of each metabolite; and 3) validation of the quantification method and the algorithm in a large and independent cohort. These steps are crucial for ensuring the clinical applicability of the eight-biomarker panel.

CRediT authorship contribution statement

Â. Carapito: Writing – original draft, Visualization, Formal analysis, Data curation, Conceptualization. **V.S. Fernandes Ferreira:** Writing – review & editing, Visualization, Formal analysis. **A.C. Silva Ferreira:** Writing – review & editing, Formal analysis. **A. Teixeira-Marques:** Writing – review & editing, Resources. **R. Henrique:** Writing – review & editing, Resources, Methodology. **C. Jerónimo:** Writing – review & editing, Resources, Methodology. **A.C.A. Roque:** Writing – review & editing, Supervision, Funding acquisition. **F. Carvalho:** Writing – review & editing, Supervision, Funding acquisition. **J. Pinto:** Writing – review & editing, Visualization, Formal analysis, Data curation, Conceptualization. **P. Guedes de Pinho:** Writing – review & editing, Visualization, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

A national patent application (official number attributed by the National Institute of Intellectual Property 120512) protecting these results has been filed by the following authors: P. Guedes de Pinho, Â. Carapito, J. Pinto, F. Carvalho, C. Jerónimo, A.C. Silva Ferreira. The authors declare no other competing interests.

Acknowledgements

This work was funded by national funds from FCT-Fundação para a Ciência e a Tecnologia, I.P., in the scope of the Research Unit on Applied Molecular Biosciences–UCIBIO (projects UIDP/04378/2020 and UIDB/04378/2020), and the Associate Laboratory Institute for Health and Bioeconomy–i4HB (project LA/P/0140/2020). Â.C. acknowledges FCT for her PhD grant with reference 2021.05844.BD and DOI identifier 10.54499/2021.05844.BD.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2025.128749.

Data availability

The data that has been used is confidential.

References

[1] H.S. Gandomani, A.A. Tarazoj, F.H. Siri, A.K. Rozveh, S. Hosseini, N.N. Borujeni, A. Mohammadian-Hafshejani, H. Salehiniya, Essentials of bladder cancer worldwide: incidence, mortality rate and risk factors, Biomed. Res. Ther 4 (2017) 1638–1655, https://doi.org/10.15419/bmrat.v4i9.371.

[2] G.V. Flores Monar, T. Reynolds, M. Gordon, D. Moon, C. Moon, Molecular markers for bladder cancer screening: an insight into bladder cancer and FDA-approved biomarkers, Int. J. Mol. Sci. 24 (2023) 14374, https://doi.org/10.3390/ijms241814374.

[3] Z. Fan, H. Shi, J. Luo, X. Guo, B. Wang, Y. Liu, J. Yu, Diagnostic and therapeutic effects of fluorescence cystoscopy and narrow-band imaging in bladder cancer: a systematic review and network meta-analysis, Int. J. Surg. 109 (2023) 3169–3177, https://doi.org/10.1097/JS9.0000000000000592.

[4] S.F. Shariat, J.A. Karam, Y. Lotan, P.I. Karakiewizc, Critical evaluation of urinary markers for bladder cancer detection and monitoring, Rev. Urol. 10 (2008) 120–135.

[5] Q. Xie, Z. Huang, Z. Zhu, X. Zheng, J. Liu, M. Zhang, Y. Zhang, Diagnostic value of urine cytology in bladder cancer. A meta-analysis, Anal Quant Cytopathol Histpathol 38 (2016) 38–44.

[6] A. Sciarra, G. Di Lascio, F. Del Giudice, P.P. Leoncini, S. Salciccia, A. Gentilucci, A. Porreca, B.I. Chung, G. Di Pierro, G.M. Busetto, E. De Berardinis, M. Maggi, Comparison of the clinical usefulness of different urinary tests for the initial detection of bladder cancer: a systematic review, Curr. Urol. 15 (2021) 22–32, https://doi.org/10.1097/CU9.0000000000000012.

[7] S.J. Galgano, S. Rais-Bahrami, K.K. Porter, C. Burgan, The role of imaging in bladder cancer diagnosis and staging, Diagnostics 10 (2020) 703, https://doi.org/10.3390/diagnostics10090703.

[8] A.G. van der Heijden, J.A. Witjes, Recurrence, progression, and Follow-Up in non-muscle-invasive bladder cancer, Eur. Urol. Suppl. 8 (2009) 556–562, https://doi.org/10.1016/j.eursup.2009.06.010.

[9] P. Gontero, A. Birtle, E. Compérat, J.L. Dominguez, F. Liedberg, P. Mariappan, A. Masson-Lecomte, A.H. Mostafid, B.W.G. Van Rhijn, T. Seisen, S.F. Shariat, E. N. Xylinas, EAU Guidelines on non-muscle-invasive Bladder Cancer (TaT1 and CIS), 2024.

[10] A. Richters, K.K.H. Aben, L.A.L.M. Kiemeney, The global burden of urinary bladder cancer: an update, World J. Urol. 38 (2020) 1895–1904, https://doi.org/10.1007/s00345-019-02984-4.

[11] P. Scilipoti, M. Moschini, R. Li, S.P. Lerner, P.C. Black, A. Necchi, M. Rouprêt, S. F. Shariat, S. Gupta, A.K. Morgans, S.P. Psutka, A.M. Kamat, The financial burden of localized and metastatic bladder cancer, Eur. Urol. 87 (2025) 536–550, https://doi.org/10.1016/j.eururo.2024.12.002.

[12] H.H. Lee, S.H. Kim, Review of non-invasive urinary biomarkers in bladder cancer, Transl. Cancer Res. 9 (2020) 6554–6564, https://doi.org/10.21037/tcr-20-1990.

[13] R. Chou, J.L. Gore, D. Buckley, R. Fu, K. Gustafson, J.C. Griffin, S. Grusing, S. Selph, Urinary biomarkers for diagnosis of bladder cancer: a systematic review and meta-analysis, Ann. Intern. Med. 163 (2015) 922–931, https://doi.org/10.7326/M15-0997.

[14] Medical Policy Manual, Urinary biomarkers for cancer screening, diagnosis, and surveillance, Regence 72 (2025) 1–17.

[15] S.S. Sugeeta, A. Sharma, K. Ng, A. Nayak, N. Vasdev, Biomarkers in bladder cancer surveillance, Front. Surg. 8 (2021) 735868, https://doi.org/10.3389/fsurg.2021.735868.

[16] S. Qiu, Y. Cai, H. Yao, C. Lin, Y. Xie, S. Tang, A. Zhang, Small molecule metabolites: discovery of biomarkers and therapeutic targets, Signal Transduct. Targeted Ther. 8 (2023) 132, https://doi.org/10.1038/s41392-023-01399-3.

[17] M. Khoubnasabjafari, M.R.A. Mogaddam, E. Rahimpour, J. Soleymani, A.A. Saei, A. Jouyban, Breathomics: review of sample collection and analysis, data modeling and clinical applications, Crit. Rev. Anal. Chem. 52 (2022) 1461–1487, https://doi.org/10.1080/10408347.2021.1889961.

[18] S. Janfaza, B. Khorsand, M. Nikkhah, J. Zahiri, Digging deeper into volatile organic compounds associated with cancer, Biol. Methods Protoc. 4 (2019) 1–11, https://doi.org/10.1093/biomethods/bpz014.

[19] T. Furuhashi, K. Toda, W. Weckwerth, Review of cancer cell volatile organic compounds: their metabolism and evolution, Front. Mol. Biosci. 11 (2025) 1499104, https://doi.org/10.3389/fmolb.2024.1499104.

[20] F. Gouzerh, J.M. Bessière, B. Ujvari, F. Thomas, A.M. Dujon, L. Dormont, Odors and cancer: current status and future directions, Biochim. Biophys. Acta Rev. Canc 1877 (2022) 188644, https://doi.org/10.1016/j.bbcan.2021.188644.

[21] C.M. Willis, S.M. Church, C.M. Guest, A. Cook, N. Mccarthy, A.J. Bransbury, M.R. T. Church, J.C.T. Church, Olfactory detection of human bladder cancer by dogs: proof of principle study, BMJ 329 (2004) 712, https://doi.org/10.1136/bmj.329.7468.712.

[22] G. Horvath, H. Andersson, G. Paulsson, Characteristic odour in the blood reveals ovarian carcinoma, BMC Cancer 10 (2010) 643, https://doi.org/10.1186/1471-2407-10-643.

[23] C. Reeve, E. Cummings, E. McLaughlin, S. Smith, S. Gadbois, An idiographic investigation of diabetic alert dogs' ability to learn from a small sample of breath samples from people with type 1 diabetes, Can. J. Diabetes 44 (2020) 37–43, https://doi.org/10.1016/j.jcjd.2019.04.020.

[24] C.Q. Gao, S.N. Wang, M.M. Wang, J.J. Li, J.J. Qiao, J.J. Huang, X.X. Zhang, Y. Q. Xiang, Q. Xu, J.L. Wang, Z.H. Liu, J.G. Wang, Z.H. Chen, P.A. Hu, Z. Song, S. J. Gu, R.X. Zhang, L.F. Lei, K. Bin Zhan, Y.T. Long, Y. Zhang, M. Ye, Z. Zhong, Y. B. Liu, C. Zhang, Z.M. He, X. Fang, J.G. Peng, C.Y. Wang, H. Xu, B.H. Xia, L. Shen, B.S. Tang, C.W. Zheng, Y.A. Li, J.F. Guo, Sensitivity of sniffer dogs for a diagnosis of parkinson's disease: a diagnostic accuracy study, Mov. Disord. 37 (2022) 1807–1816, https://doi.org/10.1002/mds.29180.
[25] H. Williams, A. Pembroke, Sniffer dogs in the melanoma clinics? Lancet 1 (1989) 734, https://doi.org/10.1016/s0140-6736(89)92257-5.
[26] U.W. Liebal, A.N.T. Phan, M. Sudhakar, K. Raman, L.M. Blank, Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (2020) 243, https://doi.org/10.3390/metabo10060243.
[27] Q. Al-Tashi, M.B. Saad, A. Muneer, R. Qureshi, S. Mirjalili, A. Sheshadri, X. Le, N. I. Vokes, J. Zhang, J. Wu, Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review, Int. J. Mol. Sci. 24 (2023) 7781, https://doi.org/10.3390/ijms.
[28] S. Badillo, B. Banfai, F. Birzele, I.I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, J.D. Zhang, An introduction to machine learning, Clin. Pharmacol. Ther. 107 (2020) 871–885, https://doi.org/10.1002/cpt.1796.
[29] C. William, C. Wangmo, A. Ranjan, Unravelling the application of machine learning in cancer biomarker discovery, Cancer Insight 2 (2023) 1–8, https://doi.org/10.58567/ci02010001.
[30] J. Pinto, Â. Carapito, F. Amaro, A.R. Lima, C. Carvalho-Maia, M.C. Martins, C. Jerónimo, R. Henrique, M. de L. Bastos, P.G. de Pinho, Discovery of volatile biomarkers for bladder cancer detection and staging through urine metabolomics, Metabolites 11 (2021) 199, https://doi.org/10.3390/metabo11040199.
[31] L. Lett, M. George, R. Slater, B. De Lacy Costello, N. Ratcliffe, M. García-Fiñana, H. Lazarowicz, C. Probert, Investigation of urinary volatile organic compounds as novel diagnostic and surveillance biomarkers of bladder cancer, Br. J. Cancer 127 (2022) 329–336, https://doi.org/10.1038/s41416-022-01785-8.
[32] K. Jobu, C. Sun, S. Yoshioka, J. Yokota, M. Onogawa, C. Kawada, K. Inoue, T. Shuin, T. Sendo, M. Miyamura, Metabolomics study on the biochemical profiles of odor elements in urine of human with bladder cancer, Biol. Pharm. Bull. 35 (2012) 639–642, https://doi.org/10.1248/bpb.35.639.
[33] H. Tyagi, E. Daulton, A.S. Bannaga, R.P. Arasaradnam, J.A. Covington, Non-invasive detection and staging of colorectal cancer using a portable electronic nose, Sensors (Basel) 21 (2021) 5440, https://doi.org/10.3390/s21165440.
[34] M. Cauchi, C.M. Weber, B.J. Bolt, P.B. Spratt, C. Bessant, D.C. Turner, C.M. Willis, L.E. Britton, C. Turner, G. Morgan, Evaluation of gas chromatography mass spectrometry and pattern recognition for the identification of bladder cancer from urine headspace, Anal. Methods 8 (2016) 4037–4046, https://doi.org/10.1039/c6ay00400h.
[35] T. Ligor, P. Adamczyk, T. Kowalkowski, I.A. Ratiu, A. Wenda-Piesik, B. Buszewski, Analysis of VOCs in urine samples directed towards of bladder cancer detection, Molecules 27 (2022) 5023, https://doi.org/10.3390/molecules27155023.
[36] H. Tyagi, E. Daulton, A.S. Bannaga, R.P. Arasaradnam, J.A. Covington, Urinary volatiles and chemical characterisation for the non-invasive detection of prostate and bladder cancers, Biosensors (Basel) 11 (2021) 437, https://doi.org/10.3390/bios11110437.
[37] A. Teixeira-Marques, S. Monteiro-Reis, D. Montezuma, C. Lourenço, M.C. Oliveira, V. Constâncio, J.P. Sequeira, C. Carvalho-Maia, R. Freitas, E.S. Martens-Uzunova, M.H. Vasconcelos, R. Henrique, C. Jerónimo, Improved recovery of urinary small extracellular vesicles by differential ultracentrifugation, Sci. Rep. 14 (2024) 12267, https://doi.org/10.1038/s41598-024-62783-9.
[38] A.R. Lima, J. Pinto, A.I. Azevedo, D. Barros-Silva, C. Jerónimo, R. Henrique, M. de Lourdes Bastos, P. Guedes de Pinho, M. Carvalho, Identification of a biomarker panel for improvement of prostate cancer diagnosis by volatile metabolic profiling of urine, Br. J. Cancer 121 (2019) 857–868, https://doi.org/10.1038/s41416-019-0585-4.
[39] I. Calejo, N. Moreira, A.M. Araújo, M. Carvalho, M. De Lourdes Bastos, P.G. De Pinho, Optimisation and validation of a HS-SPME-GC-IT/MS method for analysis of carbonyl volatile compounds as biomarkers in human urine: application in a pilot study to discriminate individuals with smoking habits, Talanta 148 (2016) 486–493, https://doi.org/10.1016/j.talanta.2015.09.070.
[40] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W. M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis: Chemical analysis working group (CAWG) metabolomics standards initiative (MSI), Metabolomics 3 (2007) 211–221, https://doi.org/10.1007/s11306-007-0082-2.
[41] M.R. Viant, I.J. Kurland, M.R. Jones, W.B. Dunn, How close are we to complete annotation of metabolomes? Curr. Opin. Chem. Biol. 36 (2017) 64–69, https://doi.org/10.1016/j.cbpa.2017.01.001.
[42] B. Quintanilla-Casas, R. Bro, J.L. Hinrich, C.L. Davie-Martin, Tutorial on PARADISe: PARAFAC2-Based deconvolution and identification system for processing GC-MS data. https://doi.org/10.21203/rs.3.pex-2143/v1, 2023.
[43] Z. Pang, Y. Lu, G. Zhou, F. Hui, L. Xu, C. Viau, A.F. Spigelman, P.E. MacDonald, D. S. Wishart, S. Li, J. Xia, MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation, Nucleic Acids Res. 52 (2024) 1–9, https://doi.org/10.1093/nar/gkae253.
[44] W. McKinney, Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, Austin, 2010, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.
[45] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[47] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[48] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.
[49] M. Waskom, Seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021, https://doi.org/10.21105/joss.03021.
[50] L.M. Antal, L.B. Iantovics, Advanced data analysis for machine learning-powered recommender systems, Procedia Comput. Sci. 246 (2024) 3957–3966, https://doi.org/10.1016/j.procs.2024.09.170.
[51] L. Berben, S.M. Sereika, S. Engberg, Effect size estimation: methods and examples, Int. J. Nurs. Stud. 49 (2012) 1039–1047, https://doi.org/10.1016/j.ijnurstu.2012.01.015.
[52] D. Wang, C. Wang, X. Pi, L. Guo, Y. Wang, M. Li, Y. Feng, Z. Lin, W. Hou, E. Li, Urinary volatile organic compounds as potential biomarkers for renal cell carcinoma, Biomed Rep 5 (2016) 68–72, https://doi.org/10.3892/br.2016.686.
[53] P. Opitz, O. Herbarth, The volatilome - investigation of volatile organic metabolites (VOM) as potential tumor markers in patients with head and neck squamous cell carcinoma (HNSCC), J. Otolaryngol. Head Neck Surg. 47 (2018) 42, https://doi.org/10.1186/s40463-018-0288-5.
[54] Q. Wen, A. Myridakis, P.R. Boshier, S. Zuffa, I. Belluomo, A.G. Parker, S.T. Chin, S. Hakim, S.R. Markar, G.B. Hanna, A complete pipeline for untargeted urinary volatolomic profiling with sorptive extraction and dual polar and nonpolar column methodologies coupled with gas chromatography time-of-flight mass spectrometry, Anal. Chem. 95 (2023) 758–765, https://doi.org/10.1021/acs.analchem.2c02873.
[55] P. Porto-Figueira, J. Pereira, W. Miekisch, J.S. Câmara, Exploring the potential of NTME/GC-MS, in the establishment of urinary volatomic profiles. Lung cancer patients as case study, Sci. Rep. 8 (2018) 13113, https://doi.org/10.1038/s41598-018-31380-y.
[56] F. Chen, C. Wang, M. Zhang, X. Zhang, Y. Liu, J. Ye, Q. Chu, Sensitive determination of endogenous hexanal and heptanal in urine by hollow-fiber liquid-phase microextraction prior to capillary electrophoresis with amperometric detection, Talanta 119 (2014) 83–89, https://doi.org/10.1016/j.talanta.2013.10.052.
[57] J.F. Liu, B.F. Yuan, Y.Q. Feng, Determination of hexanal and heptanal in human urine using magnetic solid phase extraction coupled with in-situ derivatization by high performance liquid chromatography, Talanta 136 (2015) 54–59, https://doi.org/10.1016/j.talanta.2015.01.003.
[58] D. Song, Y. Gu, L. Liang, Z. Ai, L. Zhang, H. Xu, Magnetic solid-phase extraction followed by high performance liquid chromatography for determination of hexanal and heptanal in human urine, Anal. Methods 3 (2011) 1418–1423, https://doi.org/10.1039/c1ay05102d.
[59] C.L. Silva, M. Passos, J.S. Câmara, Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry, Br. J. Cancer 105 (2011) 1894–1904, https://doi.org/10.1038/bjc.2011.437.
[60] Y. Cheng, X. Yang, X. Deng, X. Zhang, P. Li, J. Tao, C. Qin, J. Wei, Q. Lu, Metabolomics in bladder cancer: a systematic review, Int. J. Clin. Exp. Med. 8 (2015) 11052–11063.
[61] J. Huang, S. Kumar, N. Abbassi-Ghadi, P. Španěl, D. Smith, G.B. Hanna, Selected ion flow tube mass spectrometry analysis of volatile metabolites in urine headspace for the profiling of gastro-esophageal cancer, Anal. Chem. 85 (2013) 3409–3416, https://doi.org/10.1021/ac4000656.
[62] A. Pérez Antón, Á.G. Ramos, M. del Nogal Sánchez, J.L.P. Pavón, B.M. Cordero, Á. P.C. Pozas, Headspace-programmed temperature vaporization-mass spectrometry for the rapid determination of possible volatile biomarkers of lung cancer in urine, Anal. Bioanal. Chem. 408 (2016) 5239–5246, https://doi.org/10.1007/s00216-016-9618-5.
[63] K. Taunk, R. Taware, T.H. More, P. Porto-Figueira, J.A.M. Pereira, R. Mohapatra, D. Soneji, J.S. Câmara, H.A. Nagarajaram, S. Rapole, A non-invasive approach to explore the discriminatory potential of the urinary volatilome of invasive ductal carcinoma of the breast, RSC Adv. 8 (2018) 25040–25050, https://doi.org/10.1039/c8ra02083c.
[64] A. Jiménez-Pacheco, M. Salinero-Bachiller, M.C. Iribar, A. López-Luque, J.L. Mián-Ortiz, J.M. Peinado, Furan and p-xylene as candidate biomarkers for prostate cancer, Urol. Oncol.: Seminars and Original Investigations 36 (2018) 243.e21–243.e27, https://doi.org/10.1016/j.urolonc.2017.12.026.
[65] Q. Liu, Y. Fan, S. Zeng, Y. Zhao, L. Yu, L. Zhao, J. Gao, X. Zhang, Y. Zhang, Volatile organic compounds for early detection of prostate cancer from urine, Heliyon 9 (2023) e16686, https://doi.org/10.1016/j.heliyon.2023.e16686.
[66] K. Taunk, P. Porto-Figueira, J.A.M. Pereira, R. Taware, N.L. da Costa, R. Barbosa, S. Rapole, J.S. Câmara, Urinary volatomic expression pattern: paving the way for identification of potential candidate biosignatures for lung cancer, Metabolites 12 (2022), https://doi.org/10.3390/metabo12010036.
[67] C.L. Silva, R. Perestrelo, P. Silva, H. Tomás, J.S. Câmara, Volatile metabolomic signature of human breast cancer cell lines, Sci. Rep. 7 (2017) 43969, https://doi.org/10.1038/srep43969.

[68] P. Porto-Figueira, J.A.M. Pereira, J.S. Câmara, Exploring the potential of needle trap microextraction combined with chromatographic and statistical data to discriminate different types of cancer based on urinary volatomic biosignature, Anal. Chim. Acta 1023 (2018) 53–63, https://doi.org/10.1016/j.aca.2018.04.027.

[69] X. Chen, F. Xiang, X. Cao, J. Zou, B. Zhang, X. Ding, Effects of p-cresol, a uremic toxin, on cancer cells, Transl. Cancer Res. 12 (2023) 367–374, https://doi.org/10.21037/tcr-22-2042.

[70] Y.H. Hsu, H.P. Huang, H.R. Chang, The uremic toxin p-cresol promotes the invasion and migration on carcinoma cells via ras and mTOR signaling, Toxicol. Vitro 58 (2019) 126–131, https://doi.org/10.1016/j.tiv.2019.03.029.