# Enhancing Diagnostic Accuracy for Urinary Tract Disease through Explainable SHAP-Guided Feature Selection and Classification

Filipe Ferreira de Oliveira
*Labcin*
*Federal University of Espírito Santo*
Vitória, Brazil
0009-0009-9961-7329

Matheus Becali Rocha
*Labcin and PPGI*
*Federal University of Espírito Santo*
Vitória, Brazil
0009-0003-9957-5276

Renato A. Krohling
*Labcin and PPGI*
*Federal University of Espírito Santo*
Vitória, Brazil
0000-0001-8861-4274

***Abstract*—This paper proposes an explainable machine learning framework to support the diagnosis of urinary tract diseases, particularly bladder cancer, by employing SHAP (SHapley Additive exPlanations)-based feature selection to enhance model transparency and predictive performance. Six binary classification scenarios were designed to differentiate bladder cancer from other urological and oncological conditions. The models were implemented using XGBoost, LightGBM, and CatBoost algorithms, with hyperparameter optimization conducted via Optuna and class imbalance addressed through the SMOTE technique. Predictive variables were selected according to their importance values derived from SHAP analyses. The results demonstrated that SHAP-based feature selection efficiently reduced data dimensionality while maintaining or even improving performance metrics such as balanced accuracy, precision, and specificity. These findings underscore the potential of explainable artificial intelligence in medical diagnostics. The proposed methodology contributes to the development of transparent, reliable, and efficient clinical decision support systems, helping to optimize the screening and early diagnosis of urinary tract diseases, particularly bladder cancer.**

***Index Terms*—Artificial Intelligence, Explainable, Urinary Tract Diseases, SHAP, Dimensionality Reduction**

## I. Introduction

Artificial Intelligence (AI) has rapidly expanded across various domains, including the medical field [1]. However, many of these models, particularly those based on deep learning, operate as black boxes, providing high accuracy but low transparency regarding the contribution of input variables to the final outcome [2]. This lack of interpretability is especially critical in high-risk contexts such as clinical diagnosis, with bladder cancer (BC) being a prominent example. Classified as the tenth most common type of neoplasm and the thirteenth leading cause of cancer-related death worldwide [3], with more than 90% of cases belonging to the urothelial carcinoma subtype [4], BC presents both high incidence and heterogeneous clinical behavior. In this scenario, improving early diagnostic methods is essential, making the application of transparent and interpretable AI models crucial to reducing the morbidity and mortality associated with this neoplasm.

Smoking is identified as the main risk factor, being associated with approximately half of the cases and 37% of disease-related deaths [3]. Occupational exposure to carcinogenic substances represents the second most relevant factor, accounting for about 10% of cases [3], which contributes to an incidence up to four times higher in men compared to women. Lifestyle and environment-related factors also play a role, as the urothelial epithelium is continuously exposed to potentially mutagenic substances present in the urine [5]. The most frequently reported symptom is hematuria (blood in the urine), which may be macroscopic (visible) or microscopic (detected through laboratory tests), typically intermittent and painless [6], [7].

Beyond its clinical impact, bladder cancer imposes a significant financial burden on healthcare systems due to its high recurrence rate and the costs of diagnostic procedures, such as periodic cystoscopies, intravesical treatments, and more complex surgeries [4], [5]. In developed regions, where population aging is increasing, the global burden of BC is projected to rise in the coming decades, underscoring the need for improved strategies for prevention, diagnosis, and treatment [3].

In this context, the development of more accurate and interpretable models and analytical methods may play a relevant role in early detection and therapeutic management. Approaches that enhance the understanding of the determinants of bladder cancer progression have the potential to optimize clinical decision-making and reduce the associated morbidity and mortality.

Clinical and epidemiological databases often present high dimensionality, with many variables and, at the same time, a relatively small number of samples. Under these conditions, conventional machine learning algorithms may produce models with limited generalization power. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) [8] and SHAP (SHapley Additive exPlanations) [9], help mitigate this issue but do not fully address the challenge of interpretability [2].

This study proposes the application of SHAP for dimension-

ality reduction. The aim is to identify potential relationships between variables with high predictive importance and the outcome of interest. For the analysis, a dataset consisting of 1,336 clinical and laboratory samples from patients with different urinary tract diseases, collected by Tsai et al. [10] between 2017 and 2022, will be used.

Although SHAP was originally developed for feature interpretation, it can also be applied to dimensionality reduction. Kumar et al. [11] employed this technique on two datasets, demonstrating that, in addition to assigning importance to variables, it enables the simplification of complex models without compromising the reliability of predictions.

Liu et al. [12] employed machine learning models combined with SHAP for feature selection in the diagnosis of Parkinson's disease, integrating algorithms such as Deep Forest (gcForest), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF). The results indicated superior performance compared to conventional techniques, achieving accuracy above 91% and an F1-score of 0.945.

The structure of this article is organized as follows: Section II presents in detail the data used in the study, the machine learning models employed, and the experimental methodology adopted; Section III describes the experimental setup, the results obtained, and the associated discussion. Subsequently, Section IV addresses the critical analysis of the results and relevant interpretations. Finally, Section V provides the study's conclusions and indicates possible directions for future work.

## II. Material and Methods

### A. Patient data

The dataset used consists of clinical and laboratory records, including urinalysis and biochemical tests, collected at Mackay Memorial Hospital between January 2017 and February 2020 [10]. It comprises 1,336 samples distributed across five classes: 591 cases of bladder cancer, 201 cases of prostate cancer, 200 cases of kidney cancer, 200 cases of uterine cancer, and 144 cases of cystitis, as presented in Table I. The dataset contains a total of 39 variables, a considerable number that reinforces the need for dimensionality reduction.

TABLE I: Class distribution of the dataset.

| Disease | Number of samples | Missing data (%) |
|---|---|---|
| Bladder cancer | 591 | 42,27 |
| Prostate cancer | 201 | 15,04 |
| Kidney cancer | 200 | 14,97 |
| Uterus cancer | 200 | 14,97 |
| Cystitis | 144 | 10,78 |

### B. Data Preprocessing

*1) Missing Data Imputation:* The treatment of missing values is a critical step in data preprocessing, as failures in extraction or collection may generate gaps that compromise model performance [13]. Proper imputation helps reduce biases and preserve the consistency of subsequent analyses. In this study, the KNNImputer, based on the K-Nearest Neighbors (KNN) algorithm, was employed [14]. The method estimates missing values from the nearest samples in the feature space, assuming that similar instances exhibit correlated patterns [13]. The procedure follows three main steps: (i) for each instance with missing values, a set of $k$ nearest neighbors is identified according to a distance metric, such as Euclidean distance; (ii) the known values of these neighbors are used to estimate the missing value; (iii) the imputation is carried out using the mean, median, or mode, depending on the variable type and the adopted configuration.

*2) Oversampling:* Imbalanced datasets, in which some classes are underrepresented, can hinder the generalization of machine learning algorithms for those classes [15]. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) [16] was applied, which generates new synthetic samples instead of simply replicating minority instances. The method selects a minority class instance, identifies its $k$ nearest neighbors, randomly chooses one of them, and creates a new sample at a random point along the line segment connecting the original instance to the neighbor; this process is repeated until the desired class proportion is achieved [17].

*3) Dimensionality Reduction:* In high-dimensional datasets, the presence of irrelevant or redundant variables can degrade the performance of machine learning algorithms, increase computational cost, and reduce model interpretability [12], [18]. To mitigate these effects, feature selection methods such as SHapley Additive exPlanations (SHAP) [9] allow the identification and prioritization of variables with greater predictive impact. SHAP decomposes each prediction into the sum of the individual contributions of the variables, providing explainability and supporting feature selection. Although the exact calculation based on Shapley values is infeasible in high-dimensional problems, SHAP employs computational approximations, such as weighted linear regression for generic models and specific optimizations for tree-based models [19]. The SHAP value of a variable $i$ indicates its average marginal contribution to the prediction, considering all possible orders of inclusion of the variables in the model.

### C. Algorithms

*1) Extreme Gradient Boosting (XGBoost):* eXtreme Gradient Boosting (XGBoost) [20] is a tree-based Gradient Boosting algorithm, recognized for its high performance, scalability, and efficient regularization for overfitting control. The boosting technique consists of iteratively combining weak models (typically shallow decision trees) to form a robust model.

*2) Categorical Boosting (CatBoost):* Categorical Boosting (CatBoost) [21] is a Gradient Boosting algorithm based on decision trees that directly handles categorical variables, eliminating the need for manual encoding. It employs Ordered Target-Based Encoding to prevent target leakage and incorporates regularization to reduce overfitting, making it particularly efficient in problems with many categorical variables, such as recommendation systems and demand forecasting.

*3) Light Gradient Boosting Machine (LightGBM):* The Light Gradient Boosting Machine (LightGBM) [22] is a Gradi-

ent Boosting algorithm optimized for large-scale data and high dimensionality. It employs Gradient-Based One-Side Sampling (GOSS) to prioritize instances with larger gradients and reduce sample size without information loss. In addition, it adopts a leaf-wise growth strategy, which expands the leaf that yields the greatest error reduction, requiring regularization to prevent overfitting. These optimizations make training faster and more efficient, making LightGBM particularly suitable for large-scale applications such as fraud detection and recommendation systems.

### *D. Experimental Methodology*

The experimental methodology adopted was inspired by the study of Tsai et al. [10], who used clinical and laboratory data for bladder cancer prediction. Similarly, this work employs a binary classification strategy, defining multiple experimental scenarios, each designed to discriminate between a specific pair of pathological conditions. The scenarios analyzed were: (1) Bladder Cancer vs. Prostate Cancer, (2) Bladder Cancer vs. Cystitis, (3) Bladder Cancer vs. Kidney Cancer, (4) Bladder Cancer vs. Uterus Cancer, (5) Bladder Cancer vs. Others, and (6) Prostate Cancer vs. Others. In experiments 5 and 6, the goal was to identify cases of bladder or prostate cancer against the other conditions in the dataset, where the class distribution in each scenario is presented in Table II.

TABLE II: Class Distribution by Experiment

| Experiment | Groups | Number of Samples |
|---|---|---|
| 1 | Bladder Cancer | 591 |
| | Prostate Cancer | 201 |
| 2 | Bladder Cancer | 591 |
| | Cystitis | 144 |
| 3 | Bladder Cancer | 591 |
| | Kidney Cancer | 200 |
| 4 | Bladder Cancer | 591 |
| | Uterus Cancer | 200 |
| 5 | Bladder Cancer | 591 |
| | Others | 745 |
| 6 | Prostate Cancer | 201 |
| | Others | 1135 |

A pipeline was developed to integrate all preprocessing steps and the classification model, serving as the central object of optimization and training. Figure 1 illustrates its structure. Data preprocessing involved imputing missing values for numerical variables using the KNNImputer, with the number of neighbors $k$ defined through optimization with Optuna, followed by standardization with StandardScaler. For categorical variables, imputation was performed using the mode with SimpleImputer, followed by One-Hot Encoding, resulting in a total of 56 variables for the complete model. Class balancing was applied exclusively to the training data during cross-validation, using the SMOTE technique. The classification step employed machine learning algorithms, including XGBoost, CatBoost, and LightGBM. The source code is available at: https://github.com/filipefolv/shap-utd-explainability-study.

*1) Hyperparameters setting:* The hyperparameter search was conducted using Optuna, with the goal of maximizing the mean Balanced Accuracy (BACC) obtained through stratified 5-fold cross-validation. In this procedure, the dataset of each

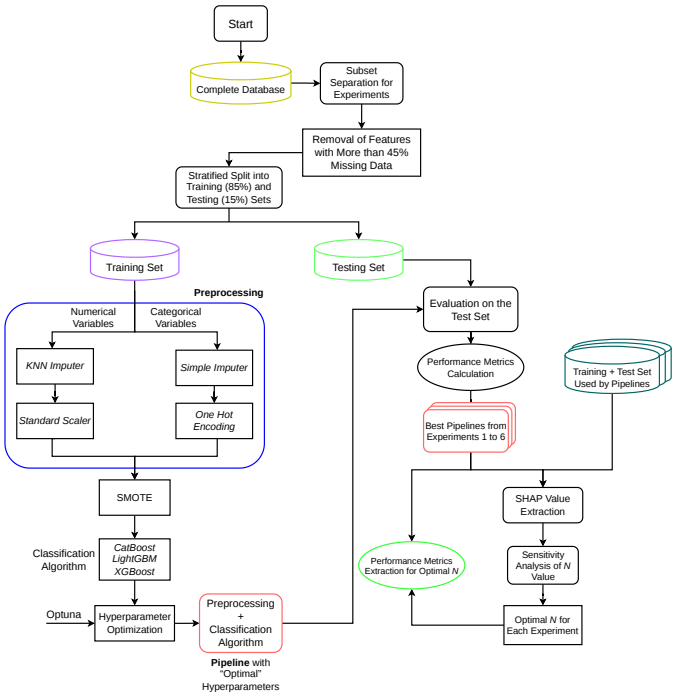



Fig. 1: General framework: the dataset is divided according to experiments and classes. The training set undergoes preprocessing (imputation, normalization, and encoding) and oversampling with SMOTE, followed by hyperparameter optimization using Optuna. With the optimal parameters, the model is trained and evaluated, generating performance metrics. Subsequently, SHAP values are used for dimensionality reduction and the selection of the most relevant features for classification.

scenario was divided into five subsets (folds) of approximately equal size, while preserving the original class distribution within each fold. In each iteration, four folds were used to train the model and the remaining fold for hyperparameter validation, repeating the process until each fold had served once as the validation set. The mean BACC across the five folds provided an estimate of model performance for the tested hyperparameter configuration. A total of 100 optimization trials per scenario and per model were performed to identify the parameter combination that maximized BACC. The hyperparameter search space is listed in Table III.

## III. Results

Six binary classification experiments were conducted, applied under two approaches: using the complete dataset and the dataset after dimensionality reduction. The experiments included comparisons between bladder cancer and different urological and oncological conditions, as well as broader scenarios involving multiple pathologies, including comparisons with the baseline [10]. For each binary scenario, a standardized computational pipeline was employed, implemented in Python using the Scikit-learn [23], Imbalanced-learn [24], and Optuna [25] libraries.

TABLE III: Hyperparameter search space for XGBoost, CatBoost, LightGBM and KNNImputer.

| Algorithm | Hyperparameter | Search space |
|---|---|---|
| XGBoost | max depth | [3 ∼ 15] |
| | number of estimators | [50 ∼ 250] |
| | learning rate | [0,01 ∼ 0,2] |
| | subsample | [0,5 ∼ 1,0] |
| | colsample by tree | [0,6 ∼ 1,0] |
| | min child weight | [1,0 ∼ 10] |
| | gamma | [0,0 ∼ 1,0] |
| | reg alpha | [0,01 ∼ 2,0] |
| | reg lambda | [0,01 ∼ 5,0] |
| | scale pos weight | [0,5 ∼ 5,0] |
| LightGBM | max depth | [3 ∼ 15] |
| | number of leaves | [20 ∼ 50] |
| | number estimators | [50 ∼ 200] |
| | learning rate | [0,01 ∼ 0,2] |
| | subsample | [0,5 ∼ 1,0] |
| | colsample by tree | [0,6 ∼ 1,0] |
| | min child samples | [5 ∼ 100] |
| | reg lambda | [0,01 ∼ 5,0] |
| | lambda l1 | [0,0 ∼ 2,0] |
| | lambda l2 | [0,0 ∼ 5,0] |
| | min split gain | [0,0 ∼ 1,0] |
| CatBoost | depth | [3 ∼ 15] |
| | learning rate | [0,01 ∼ 0,2] |
| | iterations | [50 ∼ 200] |
| | l2 leaf reg | [1 ∼ 10] |
| | bagging temperature | [0,1 ∼ 5,0] |
| | border count | [32 ∼ 255] |
| | random strength | [0,5 ∼ 2,0] |
| KNNImputer | number of neighbors | [3 ∼ 30] |

The preprocessing stage included the treatment of missing data, where a threshold of 45% was defined for the removal of variables with high rates of missingness. Variables exceeding this threshold were excluded from the analysis, as in the case of the attribute Calcium in the bladder cancer (BC) vs. uterus cancer (UC) experiment, which presented 46.4% missing values. After this cleaning step, the variable with the highest percentage of remaining missing data in the dataset had 44.2%. Following data cleaning, the dataset was initially split into training and test sets, with 85% of the data allocated for training and 15% reserved for testing. The training subset was then further partitioned using a cross-validation (CV) strategy, allocating 80% of the data for training and 20% for validation within each fold. In this study, $k = 5$ folds were employed, corresponding to the standard configuration provided by the Scikit-learn library [23].

For both approaches, three classification pipelines were constructed, corresponding to the evaluated models (XGBoost, LightGBM, and CatBoost). After hyperparameter optimization with Optuna and evaluation on the test set, the model with the highest Balanced Accuracy (BACC) was selected for each experiment. In the reduced-data approach, an additional step was performed: keeping the same train-test split, SHAP was applied to compute feature importance and assess the impact of dimensionality reduction. The number $N$ of variables with the highest SHAP values was defined as the one that yielded the best BACC, determined through a sensitivity test ranging from 2 up to the total number of available variables in each experiment. The results obtained, with and without feature reduction, are presented in Table IV, while the 20 variables with the highest SHAP values are illustrated in Figure 2.

The results indicate that dimensionality reduction had a positive impact in most experiments. In the Bladder Cancer (BC) vs. Prostate Cancer (PC) experiment, the variables with the highest SHAP values were *Urine epithelium* and *Gender* (Figure 2a). Since prostate cancer occurs exclusively in male individuals, the model's attribution of a high SHAP value to this variable is consistent. The best BACC was $93.00 \pm 1.25\%$ for the reduced model with $N = 3$ compared to $91.08 \pm 2.67\%$ for the entire model. When compared with the baseline, both the reduced and complete models achieved better accuracy (ACC), with the reduced model yielding the best performance at $95.00 \pm 1.12\%$ against $84.80\%$ from the baseline.

In the BC vs. Cystitis experiment, two features stood out: *Urine epithelium* and *A/G Ratio* (Figure 2b). A BACC of $94.24 \pm 2.43\%$ was obtained with $N = 24$, compared to $90.51 \pm 1.73\%$ for the entire model. When compared with the baseline, the reduced model achieved $92.43 \pm 2.02\%$ against $87.60\%$. It is also noteworthy that both precision and specificity reached above $97\%$. In the BC vs. Kidney Cancer (KC) experiment, the feature *Urine Occult Blood* showed a high SHAP value (Figure 2c), being a relevant marker in the diagnosis of bladder cancer [6], [7]. The BACC results were $68.17 \pm 1.43\%$ obtained with $N = 35$, compared to $65.60 \pm 1.76\%$ for the entire model. When compared with the baseline, the experiments yielded lower results in terms of ACC, Sensitivity, Specificity, and F1-Score. Notable differences were observed for ACC, which decreased from $84.50\%$ in the baseline to $70.25 \pm 2.67\%$, and for Specificity, which declined from $82.9\%$ to $64.00 \pm 4.90\%$.

In the BC vs. UC experiment, *Gender* and *Urine epithelium* again showed high SHAP values (Figure 2d), reflecting the influence of gender in diseases exclusive to the female sex. The two most important features were the same as those identified in the BC vs. PC experiment. The best BACC of $95.65 \pm 1.24\%$ was achieved by the entire model, compared with $95.00 \pm 1.39\%$ for the reduced model with $N = 14$. When compared to the baseline, the reduced model achieved $96.50 \pm 0.94\%$ ACC, whereas the baseline reached $86.90\%$ ACC.

In the BC vs. All experiment, the best BACC was achieved by the entire model, reaching $80.52 \pm 1.09\%$ compared to $78.81 \pm 0.91\%$ obtained by the reduced model with $N = 33$ features. These results again highlight the relevance of the *Urine epithelium* feature, as illustrated in Figure 2e). In the PC vs. All experiment, the entire model also yielded the best performance, achieving a BACC of $89.29 \pm 1.70\%$, whereas the reduced model with $N = 32$ features reached $85.93 \pm 1.72\%$. According to SHAP values, *Gender* was identified as the most influential variable (Figure 2f). For these final two experiments, no baseline values were available, and the entire model was used as the reference.

## IV. Discussion

In several experiments, the metrics outperformed those reported by Tsai et al. [10], particularly in terms of Balanced Accuracy (BACC), precision, and specificity. It can be observed that, in most scenarios, dimensionality reduction either

TABLE IV: Performance achieved by the models in each experiment, considering both the results with the complete dataset and after dimensionality reduction. The Baseline rows refer to the reference values from [10], while the Entire rows indicate the results obtained with the full dataset.

| Exp. | Model | Alg. | *N* | ACC(%) | BACC(%) | Prec.(%) | Sens.(%) | Spec.(%) | F1(%) |
|---|---|---|---|---|---|---|---|---|---|
| BC vs. PC | Reduced | LightGBM | 3 | **95,00 ± 1,12** | **93,00 ± 1,25** | **96,36 ± 0,64** | **97,00 ± 1,25** | **89,00 ± 2,00** | **96,67 ± 0,76** |
| | Entire | LightGBM | 57 | $92,94 \pm 1,26$ | $91,08 \pm 2,67$ | $95,74 \pm 1,88$ | $94,83 \pm 0,90$ | $87,33 \pm 5,73$ | $95,27 \pm 0,80$ |
| | Baseline [10] | LightGBM | - | 84,80 | - | 86,60 | 84,40 | 85,10 | 85,10 |
| BC vs. Cystitis | Reduced | LightGBM | 24 | **92,43 ± 2,02** | **94,24 ± 2,43** | **99,27 ± 0,89** | **91,33 ± 1,94** | **97,14 ± 3,50** | **95,13 ± 1,32** |
| | Entire | LightGBM | 57 | $90,81 \pm 2,16$ | $90,51 \pm 1,73$ | $97,37 \pm 0,86$ | $91,01 \pm 2,93$ | $90,00 \pm 3,40$ | $94,05 \pm 1,47$ |
| | Baseline [10] | LightGBM | - | 87,60 | - | 86,30 | 89,50 | 85,50 | 87,70 |
| BC vs. KC | Reduced | CatBoost | 35 | $70,25 \pm 2,67$ | **68,17 ± 1,43** | **85,82 ± 1,13** | $72,33 \pm 4,78$ | $64,00 \pm 4,90$ | $78,40 \pm 2,63$ |
| | Entire | CatBoost | 57 | $66,39 \pm 2,06$ | $65,60 \pm 1,76$ | $84,68 \pm 0,87$ | $67,19 \pm 2,40$ | $64,00 \pm 1,33$ | $74,92 \pm 1,82$ |
| | Baseline [10] | LightGBM | - | **84,50** | - | 83,00 | **86,80** | **82,90** | **84,50** |
| BC vs. UC | Reduced | LightGBM | 14 | **96,50 ± 0,94** | $95,00 \pm 1,39$ | $97,36 \pm 0,80$ | **98,00 ± 0,67** | $92,00 \pm 2,45$ | **97,67 ± 0,62** |
| | Entire | LightGBM | 57 | $96,47 \pm 1,34$ | **95,65 ± 1,24** | **97,96 ± 0,46** | $97,30 \pm 1,52$ | **94,00 ± 1,33** | $97,63 \pm 0,92$ |
| | Baseline [10] | LightGBM | - | 86,90 | - | 87,10 | 87,80 | 86,70 | 87,30 |
| BC vs. All | Reduced | LightGBM | 33 | $78,51 \pm 0,87$ | $78,81 \pm 0,91$ | $72,96 \pm 1,03$ | $81,36 \pm 1,86$ | $76,27 \pm 1,31$ | $76,92 \pm 1,02$ |
| | Entire | LightGBM | 57 | **80,40 ± 1,32** | **80,52 ± 1,09** | **76,13 ± 2,97** | **81,57 ± 2,72** | **79,46 ± 3,83** | **78,67 ± 1,01** |
| | Baseline [10] | - | - | - | - | - | - | - | - |
| PC vs. All | Reduced | XGBoost | 32 | $84,48 \pm 0,56$ | $85,93 \pm 1,72$ | $48,90 \pm 1,04$ | $88,00 \pm 4,00$ | $83,86 \pm 0,89$ | $62,84 \pm 1,42$ |
| | Entire | XGBoost | 57 | **87,86 ± 0,67** | **89,29 ± 1,70** | **55,69 ± 1,39** | **91,33 ± 3,40** | **87,25 ± 0,57** | **69,18 ± 1,82** |
| | Baseline [10] | - | - | - | - | - | - | - | - |

maintained or improved the metrics obtained with the complete dataset.

In some cases, the removal of redundant variables or those with low SHAP importance resulted in improved performance: in the BC vs. Cystitis experiment, the use of only 24 features increased BACC from $90.51 \pm 1.73\%$ to $94.24 \pm 2.43\%$, while in the BC vs. PC experiment the best performance was achieved with only 3 variables, maintaining $97.00 \pm 1.25\%$ sensitivity. This effect can be explained by the mitigation of overfitting risk, since simpler models tend to generalize better in datasets with a limited number of samples. However, in more complex scenarios or those with strong feature overlap between classes, such as the BC vs. UC experiment, dimensionality reduction did not provide significant gains.

It is noteworthy that certain variables consistently proved important across the experiments. For example, the feature *Urine Epithelium* (UL) frequently ranked among the top two features with the highest SHAP values in all experiments except the third. The presence of epithelial cells in urinary cytology, although not a consolidated biomarker for bladder cancer, may be associated with pathological alterations in the urinary tract [26]. For all scenarios involving gender-specific diseases, the feature *Gender* was correctly ranked with a higher SHAP value. In the scenario involving bladder cancer and kidney cancer, the feature *Urine Occult Blood* obtained the highest SHAP value, which is consistent with the most frequent symptom of bladder cancer, hematuria (blood in the urine) [5]. In this dataset, the manifestation was microscopic hematuria.

The application of explainable methods such as SHAP made it possible to identify factors with a potential direct relationship to diagnosis, including laboratory and demographic variables that maintained high predictive weight. The presence of epidemiologically plausible relationships reinforces the usefulness of this approach in assisting specialists with the prioritization of examinations and the formulation of diagnostic hypotheses. It is worth noting that the dataset used presents class imbalance, which, despite the application of SMOTE, may introduce biases. Another point is that although SHAP is effective in assigning importance, it does not explicitly capture dependency relationships, making it necessary to combine it with specific methods.

## V. Conclusion

This study proposed a methodology for feature selection to support the diagnosis of bladder cancer, integrating explainability through SHAP (SHapley Additive exPlanations) within a pipeline that combined preprocessing, data imputation, class balancing with SMOTE, and hyperparameter optimization via Optuna, using three machine learning algorithms (XGBoost, LightGBM, and CatBoost). Six binary classification scenarios were conducted, comparing bladder cancer with different urological and oncological conditions as well as broader pathological contexts. Feature selection based on SHAP values effectively reduced dimensionality without significant loss of performance and, in some cases, improved metrics such as balanced accuracy, precision, and specificity. Combining explainable methods are a promising strategy for clinical decision support systems, enhancing both model transparency and interpretability.

## Acknowledgment

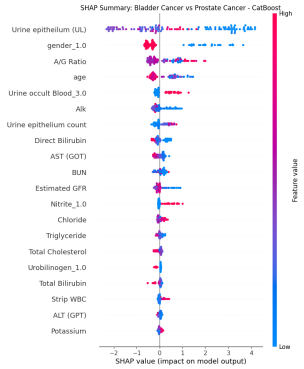
This study was financed in part by the FAPES, Brazil - Process No. 2023-CSN92, through a PhD scholarship awarded to M.B. Rocha; R.A. Krohling thanks the Brazilian research agency CNPq, Brazil – grant no. 302021/2025-6; The authors also thank the PROAP/CAPES (Portaria no. 206, 9/4/2018).

## References

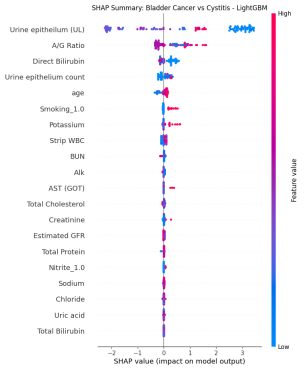
[1] T. H. Davenport, R. Ronanki *et al.*, "Artificial intelligence for the real world," *Harvard business review*, vol. 96, no. 1, pp. 108–116, 2018.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206 – 215, 2018.

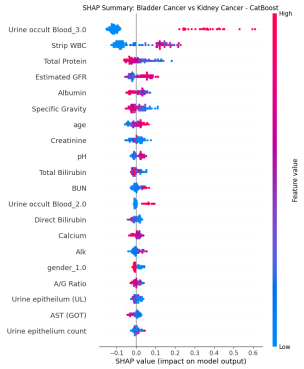
[3] Y. Zhang, H. Rumgay, M. Li, H. Yu, H. Pan, and J. Ni, "The global landscape of bladder cancer incidence and mortality in 2020 and projections to 2040," *Journal of Global Health*, vol. 13, 09 2023.



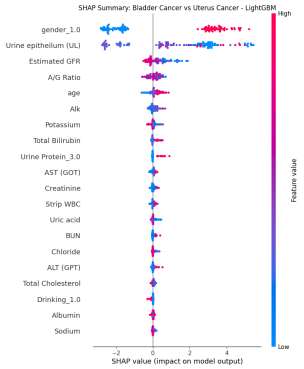

(a) Bladder Cancer vs. Prostate Cancer



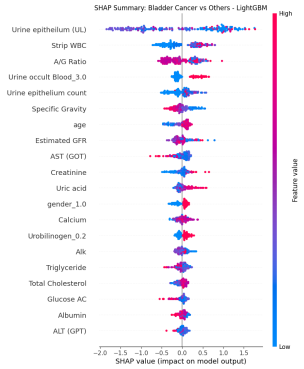

(b) Bladder Cancer vs. Cystitis



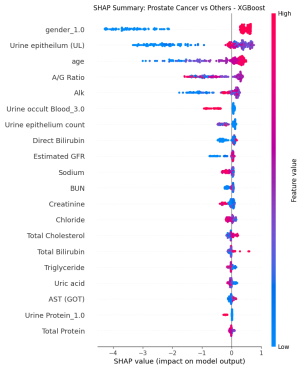

(c) Bladder Cancer vs. Kidney Cancer




(d) Bladder Cancer vs. Uterus Cancer




(e) Bladder Cancer vs. Others




(f) Prostate Cancer vs. Others

Fig. 2: SHAP values for the top 20 features of the dataset across all scenarios.


[4] L. van Hoogstraten, A. Vrieling, A. Heijden, M. Kogevinas, A. Richters, and L. Kiemeney, "Global trends in the epidemiology of bladder cancer: challenges for public health and clinical practice," *Nature Reviews Clinical Oncology*, vol. 20, pp. 287–304, 03 2023.

[5] S. A. Halaseh, S. Halaseh, Y. Alali, M. E. Ashour, and M. J. Alharayzah, "A Review of the Etiology and Epidemiology of Bladder Cancer: All You Need To Know," *Curēus (Palo Alto, CA)*, vol. 14, no. 7, pp. e27 330–, 2022.

[6] S. J. Price, E. A. Shephard, S. A. Stapley, K. Barraclough, and W. T. Hamilton, "Non-visible versus visible haematuria and bladder cancer risk: a study of electronic records in primary care," *British Journal of General Practice*, vol. 64, no. 626, pp. e584–e589, 2014.

[7] G. Dulku, A. Shivananda, A. Chakera, R. Mendelson, and D. Hayne, "Painless Visible Haematuria in Adults: An Algorithmic Approach Guiding Management," *Cureus*, vol. 11, no. 11, Nov. 2019.

[8] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space ," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[10] I.-J. Tsai, W.-C. Shen, C.-L. Lee, H.-D. Wang, and C.-Y. Lin, "Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data," *Diagnostics (Basel, Switzerland)*, vol. 12, p. 203, 01 2022.

[11] C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun, and T. Anjali, "Dimensionality Reduction based on SHAP Analysis: A Simple and Trustworthy Approach," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 558–560.

[12] Y. Liu, Z. Liu, X. Luo, and H. Zhao, "Diagnosis of Parkinson's disease based on SHAP value feature selection," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 3, pp. 856–869, 2022.

[13] T. Aljrees, "Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning," *PLOS ONE*, vol. 19, no. 1, pp. 1–24, 01 2024.

[14] G. Batista and M.-C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method," vol. 30, 01 2002, pp. 251–260.

[15] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002.

[17] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer, 2018, vol. 10.

[18] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 340–347.

[19] W. E. Marcílio-Jr and D. M. Eler, "Explaining dimensionality reduction results using Shapley values," *Expert Systems with Applications*, vol. 178, p. 115020, 2021.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6639–6649.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157.

[23] 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, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2019, p. 2623–2631.

[26] Y. Lotan and C. G. Roehrborn, "Sensitivity and specificity of commonly available bladder tumor markers versus cytology: results of a comprehensive literature review and meta-analyses," *Urology*, vol. 61, no. 1, pp. 109–118, 2003.