International Urology and Nephrology
https://doi.org/10.1007/s11255-026-05342-7

UROLOGY – ORIGINAL PAPER

Prediction of pediatric lower urinary tract dysfunction using uroflowmetry data and machine learning with explainable AI techniques

Nida Dinçel[1] · Mustafa Büyükkeçeci[2] · Caner Kivanc Hekimoglu[3] · Dersan Onur[4] · Sermet Mir[5] · Ilhan Sofuoglu[6] · Derya Ozmen[1] · Ezgi Kirmizitas[1] · Deniz Cetinkaya[7]

Received: 3 July 2026 / Accepted: 6 August 2026
© The Author(s) 2026

Abstract

Purpose To develop and evaluate a machine learning-based approach for the interpretation of uroflowmetry in pediatric lower urinary tract dysfunction, aiming to reduce interobserver variability and improve diagnostic consistency while maintaining clinical interpretability.

Methods In this single-center retrospective study, 2663 pediatric uroflowmetry records (age 2–18 years, July 2019-March 2024) were analyzed. LUTD subtypes were assigned to the International Children's Continence Society terminology. Two settings were evaluated: three-class (overactive bladder (OAB), dysfunctional voiding (DV), normal (N)) and four-class (adding DV-OAB). Eight predictors (four numerical, four categorical) were used. Seven ML models (Decision Trees, Naive Bayes, Support Vector Machine, Efficiently Trained Linear Classifiers (Logistic Regression), Gaussian Kernel, Ensemble Classifiers, and Artificial Neural Networks) were trained with Bayesian hyperparameter optimization using stratified tenfold cross-validation on a 90–10% train-test split. Weighted accuracy, precision, recall, and F1-score were reported. Feature importance was quantified with Shapley Additive Explanations (SHAP). Reporting followed TRIPOD-AI.

Results In the three-class setting, the ensemble classifier achieved the best test performance (accuracy 86.09%, F1 86.03%) and all models exceeded 82% test accuracy. In the four-class setting, test accuracy dropped to 72.56–78.95%, and most models failed to predict the DV-OAB class owing to its low prevalence (2.55%). SHAP identified bladder capacity as the dominant feature for OAB, pelvic floor activity for DV, and residual volume as a secondary DV feature. Sex showed a limited predictive contribution compared with the other clinical predictors.

Conclusions Explainable ML enables interpretable classification of pediatric LUTD subtypes from uroflowmetry data across a wide age range, including children under 5 years. External multicenter validation is required before clinical deployment.

Keywords Pediatric · Uroflowmetry · Lower urinary tract dysfunction · Explainable artificial intelligence · Machine learning algorithms

✉ Deniz Cetinkaya
dcetinkaya@bournemouth.ac.uk

[1] University of Health Sciences, Dr. Behcet Uz Children's Hospital, Izmir, Türkiye

[2] Independent Researcher, Izmir, Türkiye

[3] Department of Computer Engineering, Yaşar University, Izmir, Türkiye

[4] University of Health Sciences, Izmir Tepecik Training and Research Hospital, Izmir, Türkiye

[5] Ravensbourne University London, London, UK

[6] Department of Software Engineering, Yaşar University, Izmir, Türkiye

[7] School of Computing and Engineering, Bournemouth University, Poole House P324, Talbot Campus, Fern Barrow, Poole BH12 5BB, UK

Published online: 18 August 2026

Introduction

Pediatric lower urinary tract dysfunction (LUTD) represents a common, heterogeneous, and complex clinical spectrum in childhood that significantly impairs quality of life and, if improperly managed, can lead to upper urinary tract damage and progressive loss of renal function [1, 2]. LUTD encompasses a wide spectrum of conditions, such as overactive bladder, dysfunctional voiding, or bladder underactivity, each with distinct clinical presentations and management strategies, making accurate subtype classification a critical determinant of treatment outcomes. In the first-line evaluation of these pathologies, uroflowmetry (UF) is widely considered an essential screening tool owing to its noninvasive nature, cost-effectiveness, and ease of application [1, 3].

However, interpreting UF data and voiding curves heavily relies on the clinician's experience; studies report significant interobserver variability even among experienced specialists [3–5]. This high degree of subjectivity can lead to clinically significant consequences, such as misclassification, unnecessary advanced investigations, and delayed appropriate treatment.

To address these challenges in diagnostic consistency, computational approaches that reduce reliance on subjective interpretation have gained increasing attention. In recent years, artificial intelligence (AI) and machine learning (ML) methods have demonstrated the potential to improve diagnostic accuracy and support clinical decision-making processes in many medical fields, including pediatric nephrology and urology [6]. Within urology, the majority of published ML work has focused on the adult population or on invasive urodynamic studies, and promising results in automated UF interpretation have largely emerged from adult cohorts [7]. In pediatric urology more broadly, AI applications have grown rapidly, with notable progress in hydronephrosis grading and spina bifida-related bladder dysfunction, yet these advances have predominantly relied on imaging data or invasive urodynamic recordings rather than noninvasive UF parameters [8–14]. Research applying ML specifically to noninvasive pediatric UF data for LUTD subtype classification therefore remains limited. This situation suggests that the potential of pediatric UF in clinical decision-making processes has not yet been fully realized. This limitation is attributable to several converging factors including the lack of standardized data collection and reporting protocols across institutions, and the technical challenges of obtaining reliable UF recordings in younger children. Together, these structural limitations have substantially constrained both the quantity and quality of training data available for ML model development in this domain.

Beyond data scarcity, a further barrier to clinical adoption is the black-box nature of many high-performing ML models, which significantly restricts their integration into clinical practice. Models intended for use in clinical decision support systems must not only provide high accuracy but also transparently and comprehensibly explain the underlying determinants of their decision-making processes to clinicians [6]. Otherwise, models lacking explainability undermine clinical trust and fail to gain acceptance in real-world applications.

This study proposes a machine learning model capable of predicting and classifying LUTD subtypes in pediatric patients with high accuracy based on UF data and provides systematic evidence on UF-based classification across a wide age range, including children younger than 5 years which remains underrepresented in the current literature. To ensure that the resulting model meets the transparency requirements indispensable for clinical adoption, we further apply explainable artificial intelligence (XAI) methods to analyze the contribution of individual UF parameters to the output.

Methods

Study design and setting

This study was conducted as a single-center, retrospective, observational, cross-sectional study. A retrospective design was adopted given the availability of a sufficiently large and longitudinally maintained clinical dataset. The study was reported in accordance with the TRIPOD-AI guidelines [15].

Participants, data sources, and measurements

This retrospective study utilized a comprehensive dataset comprising 2663 pediatric UF records collected between July 2019 and March 2024. Data were systematically acquired from the electronic health records (EHR) of the Pediatric Nephrology Clinic at the University of Health Sciences, Izmir Dr. Behçet Uz Children's Hospital. During the data collection phase, the ground-truth diagnoses for the machine learning models were extracted retrospectively from the clinical evaluation notes and UF reports provided by two experienced pediatric nephrologists (authors N.D. and D.Ozmen). Each uroflowmetry session was independently evaluated by both pediatric nephrologists (N.D. and D.Ozmen), who assigned the diagnostic category as well as the categorical predictor variables derived from clinical interpretation (bladder capacity classification, voiding pattern, and pelvic floor activity) based on the same clinical assessment. Diagnostic categories were assigned by clinical judgment applying ICCS terminology. Disagreements between the two evaluators, whether concerning

the diagnostic label or the categorical predictor variables, were resolved through joint discussion until consensus was reached.

The study population included pediatric patients aged 2–18 years who presented with LUTS and were capable of complying with voiding instructions. To ensure the high quality and reliability of the data used for model training, UF recordings with significant artifacts were excluded from the dataset. The extracted data encompassed patient demographics, standard numerical UF parameters (e.g., the maximum urinary flow rate (Qmax), voided volume, post-void residual (PVR) measurements, bladder capacity) and unstructured clinical evaluation notes. Voided volume less than 50% of expected bladder capacity (EBC) or that of the maximum voided volume measured on the frequency/volume chart as well as bladder diary over 48 h were considered as inadequate and excluded.

The final analytical dataset was compiled from three sequential data extractions from the same EHR system, covering different time periods. After pooling, duplicate records arising from overlapping extraction windows were identified and removed. One record was excluded due to a physiologically implausible Qmax value. Continuous variables (Qmax, voided volume, and PVR) were further reviewed for physiological plausibility, and values deemed inconsistent were corrected on the basis of the original clinical records. The resulting dataset comprised 2663 unique patient records. Each record corresponds to a distinct uroflowmetry session to the dataset; therefore, the training, validation, and testing partitions are inherently patient-level disjoint.

Patients were classified according to International Children's Continence Society (ICCS) standardization terminology into the following categories [1]: Normal (N), overactive bladder (OAB), dysfunctional voiding (DV), and DV-OAB which patients exhibit simultaneous features of both OAB and DV.

Table 1 provides the details of the variables. For model training, two classification settings were used: a four-class setting (OAB, DV, DV-OAB, N) preserving the original diagnostic granularity and a three-class setting (OAB, DV, N), where DV-OAB cases were reclassified as OAB. This reclassification was based on two considerations: (i) the marked data scarcity of the DV-OAB subgroup (2.55% of the cohort), which substantially impaired model performance and reliable class prediction in the four-class setting, and (ii) the clinical consideration that OAB symptomatology represents an essential diagnostic component of the DV-OAB presentation, whereas DV features may co-occur without being independently required for diagnosis, supporting closer clinical proximity of DV-OAB to the OAB phenotype.

Data analysis

The experiments were implemented using MATLAB R2025a and the Statistics and Machine Learning Toolbox version 25.2 (The MathWorks Inc., Natick, MA, USA).

Table 1 List of variables

Variable	Description
Age	Patient's age in years (numeric). Bladder capacity and voiding patterns change dynamically with age, making it critical for defining personalized normal ranges
Sex	Patient's sex (binary: male/female). Anatomy and dysfunction patterns may differ by sex, and it serves as a fundamental distinguishing feature for classification
Voided volume	Total urine volume voided in tests, in mL (numeric). It is important for test validity, as low volumes can misleadingly reduce flow rates. It is also strongly correlated with QMax and useful for normalization
QMax	Maximum flow rate achieved during the test, in mL/s (numeric). It is among the most useful criteria in LUTD diagnosis and is expected to be one of the strongest predictors
Bladder capacity	The standard formula for calculating EBC = age (years) × 30 + 30 (ml) was used Expert interpretation of voided volume relative to age-expected values (categoric: normal, reduced, increased, or Approx. 2/3 of expected). It reflects the bladder's storage function
Voiding pattern	Expert classification of UF curve morphology (categoric: Plateau-shaped, Bell-shaped, Staccato-shaped, Interrupted-shaped, or Tower-shaped). It forms the target subcategories in supervised learning models
Pelvic floor	Expert interpretation of pelvic floor activity (categoric: Increased PFM tone, normal, decreased PFM tone). It represents a key mechanism underlying dysfunctional voiding
Residual	Post-void residual urine volume, in mL (numeric). High values indicate inadequate bladder emptying and are expected to have high importance in diagnosis classification
Diagnosis	Clinical diagnosis label (categoric: Normal, DV, OAB, or DV-OAB). This is the target variable that supervised learning models will predict

LUTD Lower urinary tract dysfunction, *UF* Uroflowmetry, *EBC* Expected bladder capacity, *OAB* Overactive bladder, *DV* Dysfunctional voiding, *DV-OAB* which patients exhibiting simultaneous features of both OAB and DV, *PFM* Pelvic Floor Muscles, *QMax* Maximum flow rate

The dataset consists of 2663 UF records described by four numerical and four categorical attributes and contains no missing values. In this study, two classification settings are considered (see Table 2). In the first setting, the dataset is labeled into three-classes: OAB (26.8%, 714 records). DV (42.4%, 1130 records), and N (30.8%, 819 records). In the second setting, classification is performed using four-classes: OAB (24.3%, 646 records), DV (42.4%, 1130 records), DV-OAB (2.55%, 68 records), and N (30.8%, 819 records). In both cases, the dataset is partitioned into a training set and a testing set, comprising 90% (2397 records) and 10% (266 records) of the total data, respectively. This 90:10 ratio was chosen to ensure high training data volume, which is essential for model accuracy and minority-class learning, without compromising evaluation reliability. The training set is utilized for model development, while the remaining samples are reserved for performance evaluation. The class distributions show different levels of imbalance in the two classification settings. In the three-class case, the data are fairly balanced. However, in the four-class case, the dataset becomes highly imbalanced, because the DV-OAB class represents only 2.55% of the records. This strong imbalance can make it difficult for models to correctly learn and predict the DV-OAB class. Therefore, a stratified tenfold cross-validation procedure is applied in both cases to the training set to ensure the preservation of class distributions across all folds. Consequently, this approach yields a more robust and reliable evaluation of model performance across varying data partitions.

Machine learning model development

Model performance evaluation

In this study, seven distinct classification models, including Decision Trees (DT), Naive Bayes (NB), Support Vector Machine (SVM), Efficiently Trained Linear Classifiers (ETLC), Gaussian Kernel (GK), Ensemble Classifiers, and Artificial Neural Networks (ANN), were trained using Bayesian optimization for hyperparameter tuning within the stratified tenfold cross-validation framework applied to the training set. The independent test set remained untouched during model development and was used only for final performance evaluation. Model performance is systematically evaluated via accuracy, precision, recall, and F1-score [16], ensuring a comprehensive assessment of predictive power.

In multiclass classification problems, performance metrics, such as precision, recall, and F1-score, are first computed for each class individually and subsequently aggregated to obtain a single overall performance measure. This aggregation can be performed using micro, macro, or weighted averaging strategies. Macro-averaging assigns equal importance to each class by computing the unweighted mean of class-specific metrics, whereas micro-averaging aggregates the contributions of all instances across classes to produce a global evaluation. Weighted averaging, in contrast, accounts for class imbalance by weighting each class-specific metric according to its relative support. In this study, the weighted averages of precision, recall, and F1-score are reported.

Explainable artificial intelligence (XAI)

Feature importance is quantified using Shapley Additive Explanations (SHAP) [17], a framework grounded in cooperative game theory. Shapley values assign an importance score to each predictor (feature) by calculating its average marginal contribution to the model's output across all possible feature subsets. By evaluating the contribution of each feature in the context of all possible combinations, SHAP provides a mathematical measure of global and local feature importance, allowing for the identification of the primary features behind the model's classifications.

Table 2 Dataset properties

UF records	Features	Numerical and Categorical Features	Missing Values	Classes	Class Distributions		
2663	8	4—4	No	3	OAB	714	26.8%
					DV	1130	42.4%
					N	819	30.8%
2663	8	4—4	No	4	OAB	646	24.3%
					DV	1130	42.4%
					DV-OAB	68	2.55%
					N	819	30.8%

N normal, *OAB* overactive bladder, *DV* dysfunctional voiding, *DV-OAB* simultaneous features of both OAB and DV, *Qmax* maximum flow rate

Ethics

Ethical approval was obtained from the University of Health Sciences İzmir Dr. Behçet Uz Children's Hospital Ethics Committee (279/2026/04–14).

Results

In the three-class setting experiment, all models, with the exception of GK on the validation set, achieved classification performance exceeding 80%. The Ensemble model achieved the highest scores across all metrics in both the validation and test datasets (highlighted in bold in Table 3). The ANN and ETLC (Logistic Regression) models also performed very well in both sets. The GK model showed the weakest performance in the validation set (78.93% accuracy), while the DT and NB models yielded the lowest accuracy on the test set (82.33%). As can be seen from Table 4, the results of the four-class setting indicate that the ensemble model outperformed the individual classifiers on the validation dataset (highlighted in bold in Table 4). However, on the test dataset, the DT classifier achieved superior performance relative to the other models. Across both the validation and test datasets, the performance scores of all classifiers ranged approximately between 71 and 80%. In cases where a model failed to predict any instances of a specific class, certain performance metrics could not be computed. For example, the precision score of the ANN classifier was undefined under these conditions. In Table 4, such undefined (NaN) values are denoted by hyphens. This issue arises from the models' inability to predict the DV-OAB class, consistent with the limited availability of samples for this class in the dataset. Additional detailed results for the remaining models in the four-class setting are provided in the Supplementary Material.

In the classification experiments, feature (predictor) importance was evaluated using Shapley values. While seven models were evaluated in the classification experiments, the four best-performing models from the three-class setting were selected for the Shapley-based analysis to ensure a focused interpretation of feature importance. These models also remained among the top-performing models in the four-class setting.

Figure 1 presents the feature importance of the models used in the three-class setting. As shown, the models

Table 3 Classification results for 3-class setting

Model Type	Validation				Test			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
DT	80.85	81.05	80.85	80.90	82.33	82.56	82.33	82.30
NB	81.02	81.49	81.02	81.06	82.33	82.33	82.33	82.31
SVM	80.43	80.95	80.43	80.38	83.83	83.98	83.83	83.61
ETLC	80.52	81.19	80.52	80.55	84.96	85.16	84.96	84.95
GK	78.93	79.31	78.93	78.98	82.71	82.65	82.71	82.55
Ensemble	82.06	82.50	82.06	82.05	86.09	86.24	86.09	86.03
ANN	80.81	81.51	80.81	80.85	85.34	85.54	85.34	85.30

DT decision tree, *NB* Naive Bayes, *SVM* support vector machine, *ETLC* ensemble trained linear classifier, *GK* Gaussian Kernel, *Ensemble* Ensemble Learning (Multiple model combination), *ANN* artificial neural network

Table 4 Classification results for 4-class setting

Model Type	Validation				Test			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
DT	79.68	77.97	79.68	78.77	78.95	–	78.94	77.97
NB	79.39	78.80	79.39	78.62	76.69	77.99	76.69	76.30
SVM	79.35	79.20	79.35	78.38	74.06	–	74.06	72.72
ETLC	79.47	–	79.47	78.57	75.19	–	75.19	74.35
GK	77.47	–	77.47	76.62	72.56	–	72.56	71.55
Ensemble	80.23	–	80.23	79.30	78.20	–	78.20	77.25
ANN	80.18	–	80.18	79.25	75.94	–	75.94	75.03

DT decision tree, *NB* Naive Bayes, *SVM* support vector machine, *ETLC* ensemble trained linear classifier, *GK* Gaussian Kernel, *Ensemble* Ensemble Learning (Multiple model combination), *ANN* artificial neural network

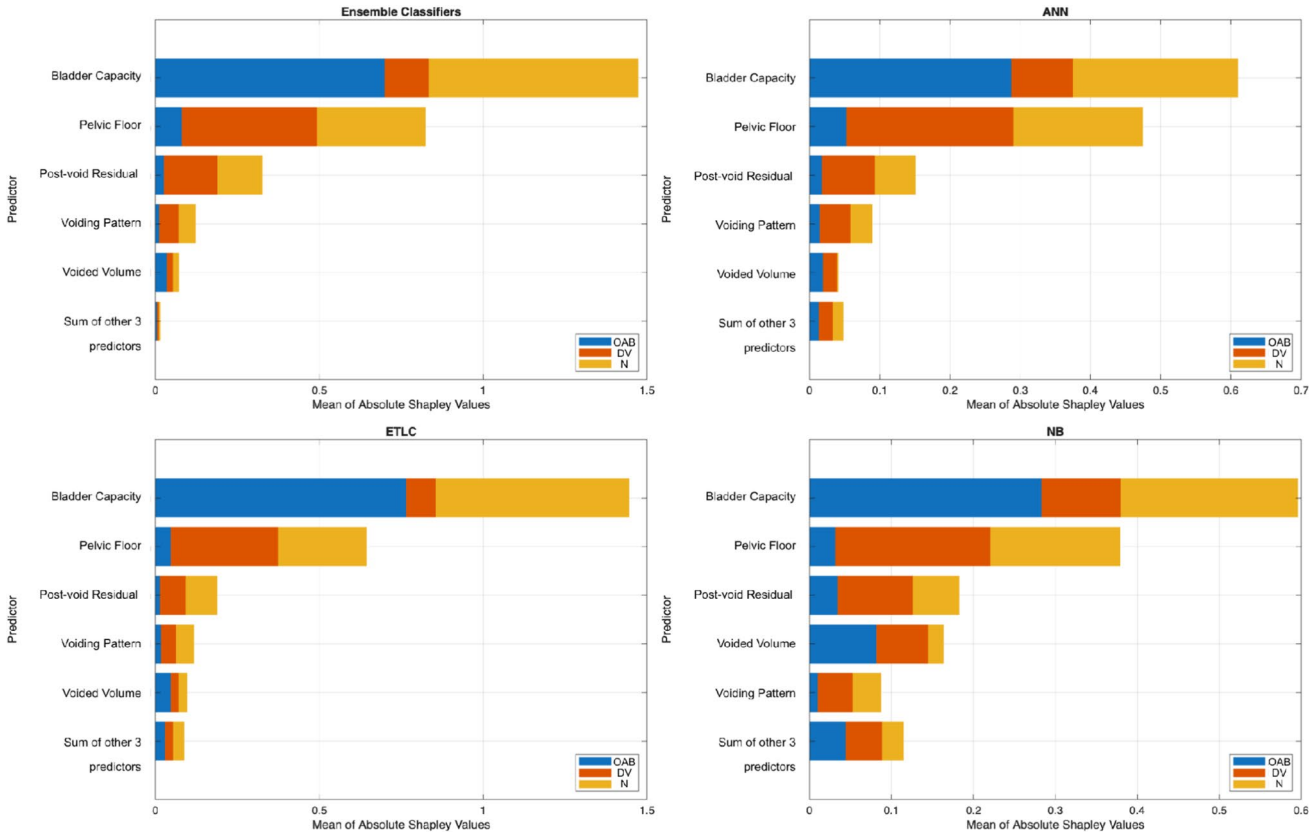


Fig. 1 3-Class Setting: mean of absolute SHAP values of NB, ETLC, Ensemble Classifiers, ANN models. The subplots present the prediction interpretations for Ensemble Classifiers (top left), Artificial Neural Network (ANN, top right), ETLC (bottom left), and Naive Bayes (NB, bottom right). The stacked horizontal bars indicate the proportional contribution of each feature toward the three target classes: OAB (blue), DV (orange), and N (yellow). *OAB* overactive bladder, *DV* dysfunctional voiding, *N* normal, *ETLC* efficiently trained linear classifiers (logistic regression), *ANN* artificial neural network, *NB* Naive Bayes

Table 5 SHAP values for ensemble classifiers model (3-class setting)

Predictors	Classes		
	OAB	DV	N
Age	0.00312	0.00476	0.00266
Sex	0	0	0
Bladder Capacity	0.69982	0.13485	0.63922
Voiding Pattern	0.01175	0.05890	0.05190
Pelvic Floor Activity	0.08083	0.41229	0.33145
Qmax	0.00107	0.00191	0.00169
Voided Volume	0.03477	0.01905	0.01784
Residual Volume	0.02612	0.16280	0.13864

N normal, *OAB* overactive bladder, *DV* dysfunctional voiding, *Qmax* maximum flow rate

generally exhibit similar results, with the effect of each feature on the corresponding diagnosis represented by the magnitude of the bars. To better quantify these differences, Table 5 presents the feature importance values for the Ensemble model only. The other models exhibited similar trends, and the Ensemble model was selected here for representational purposes. Additional detailed results for the remaining models in the three-class setting are provided in the Supplementary Material.

The feature importance for the three-class setting is visually presented as a stacked bar chart in Fig. 1, with Table 5 providing the corresponding numerical Shapley values for precise reference. Based on our analysis, the most influential features for the OAB class are bladder capacity, pelvic floor activity, and voided volume. On the other hand, the most prominent features are pelvic floor activity, residual volume, and bladder capacity for the DV class. For the N class, unlike OAB, residual volume shows greater importance instead of voided volume, with bladder capacity and pelvic floor activity also remaining significant.

Figure 2 displays the stacked bar chart visualization of the Shapley values for the four-class setting, while Table 6 details the corresponding numerical values. In this experiment, similar patterns to the three-class setting are observed for the OAB, DV, and N classes, with feature importance rankings remaining consistent. The additional

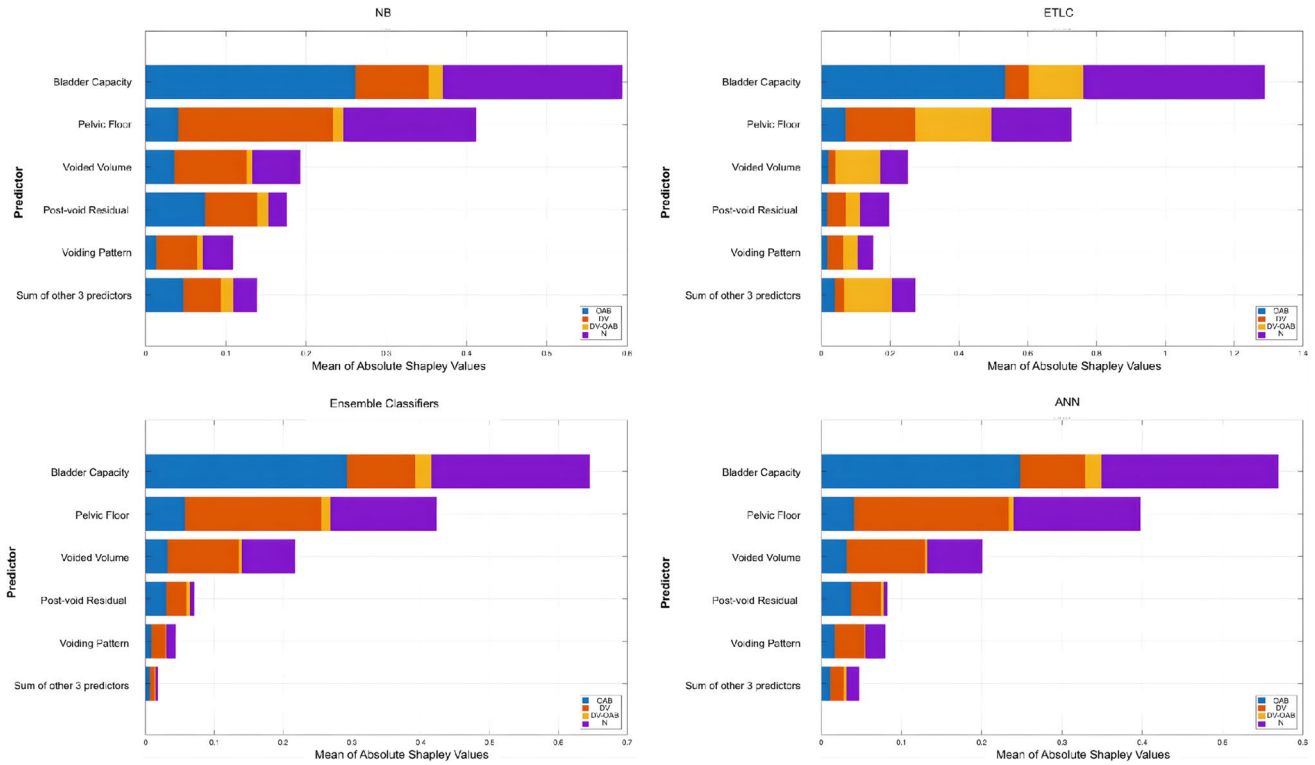


Fig. 2 4-Class setting: mean of absolute SHAP values of NB, ETLC, Ensemble Classifiers, ANN models. The subplots represent the prediction interpretations for Naive Bayes (top left), ETLC (top right), Ensemble Classifiers (bottom left), and ANN (bottom right) models. The stacked bars indicate the proportional contribution of each feature toward the four distinct target classes: OAB (blue), DV (orange), DV-OAB (yellow), and N (purple). *OAB* overactive bladder, *DV* dysfunctional voiding, *DV-OAB* which patients exhibiting simultaneous features of both OAB and DV, *N* normal, *ETLC* efficiently trained linear classifiers (logistic regression), *ANN* artificial neural network, *NB* Naive Bayes

Table 6 SHAP values for ensemble classifiers model (4-class setting)

Predictors	Classes			
	OAB	DV	DV-OAB	N
Age	0.00391	0.00494	0.00087	0.00172
Sex	0	0	0	0
Bladder Capacity	0.29241	0.09949	0.02352	0.23056
Voiding Pattern	0.00837	0.02029	0.00076	0.01460
Pelvic Floor Activity	0.05752	0.19756	0.01395	0.15399
Qmax	0.00230	0.00200	0.00030	0.00146
Voided Volume	0.02967	0.02938	0.00530	0.00652
Residual Volume	0.03205	0.10314	0.00455	0.07724

N normal, *OAB* overactive bladder, *DV* dysfunctional voiding, *DV-OAB* which patients exhibiting simultaneous features of both OAB and DV, *Qmax* maximum flow rate

DV-OAB class, as previously defined, represents patients presenting both overactive bladder symptoms and dysfunctional voiding signs. Accordingly, in this class, bladder capacity and pelvic floor activity emerge as the most influential features, corresponding to the dominant features of the OAB and DV classes, respectively.

These experiments highlight two key outcomes. First, the AI model's predictions are largely consistent with the expert-assigned clinical classifications, demonstrating the strong internal classification performance and the ability of machine learning models to identify patterns associated with uroflowmetry-based diagnostic categories. Second, the results indicate the potential of AI models for future, more comprehensive and complex studies, suggesting that AI could play a more active role in the clinical diagnostic process.

Discussion

In this study, we developed and compared seven machine learning models for the classification of LUTD subtypes in a cohort of 2663 pediatric patients based on uroflowmetry data and associated clinical parameters. In the three-class setting (OAB, DV, Normal), the Ensemble classifier achieved the highest overall accuracy (86.09%), precision (86.24%), and

F1-score (86.03%) on the test set, outperforming all individual classifiers including ANN (85.34%) and ETLC (84.96%). In the four-class setting, where the DV-OAB subtype was preserved as a distinct category, model performance declined substantially across all classifiers (72.56–78.95%), and most models failed to predict any DV-OAB cases, yielding undefined precision values. SHAP-based explainability analysis revealed that bladder capacity was the dominant predictor for OAB classification, pelvic floor activity for DV, and residual volume emerged as a secondary but clinically relevant feature for DV. In contrast, sex demonstrated a limited contribution to model predictions, with relatively low SHAP values compared with the other clinical predictors. To our knowledge, this represents the largest pediatric uroflowmetry dataset analyzed with explainable machine learning methods for LUTD subtype classification. Globally, this also represents one of the earliest applications of XAI to UF data specifically for diagnostic decision support in pediatric nephrology, extending beyond pattern classification approaches.

Despite ongoing efforts toward standardization, interpretation of UF in daily clinical practice still largely depends on the clinician's individual experience and subjective assessment [18]. Previous studies have demonstrated considerable inter-rater variability in the evaluation of UF curves and associated parameters, even among experts with similar training backgrounds [4, 19, 20]. This variability becomes more pronounced in atypical flow patterns, where the same UF curve may be classified differently by different clinicians, leading to reduced diagnostic consistency and potential variability in clinical decision-making [5]. Our findings should therefore be interpreted not only as a contribution to pediatric UF data, but also as evidence supporting the need for more objective, reproducible, and standardized evaluation tools [18].

In this regard, ML approaches have emerged as promising solutions for adult and pediatric UFs, with prior studies reporting high classification accuracies, particularly with algorithms such as XGBoost and deep learning models with accuracy rates of up to 85% and over 94%, respectively [21–24]. However, a major limitation of these approaches is their "black box" nature, which restricts their interpretability and limits clinical trust and adoption [25].

In addition to methodological considerations, the clinical applicability of UF across different age groups deserves attention. ICCS recommendations do not define a strict lower age limit for UF; however, for reliable results, the test is ideally performed in children aged ≥ 5 years and may be applied in appropriately selected and cooperative patients as young as 4 years [1]. Furthermore, in highly cooperative children who achieve early toilet training, UF has also been reported to be feasible in those younger than 4 years. In our cohort, a substantial proportion of patients fell below this threshold (2 years: n = 4, 3 years: n = 39, 4 years: n = 145), which is noteworthy in terms of systematically evaluating UF data in younger children. Given the scarcity of published data in this age group, these observations represent a notable contribution to existing literature. Additionally, the inclusion of younger children in our cohort provides novel insights into UF characteristics in early childhood, a period that remains underrepresented in the literature. Early identification of lower urinary tract dysfunction has been linked to improved long-term renal and bladder outcomes [2]. By demonstrating that ML-based classification is feasible across a wide age range, including children under 5 years, our findings support the extension of standardized UF evaluation to younger populations where timely intervention is critical.

One of the key strengths of our study is the application of XAI methods to evaluate high-performance classification models. This approach transparently demonstrates which UF features and clinical parameters influence the models' decisions, while maintaining accurate classification of UF patterns. This interpretability is crucial in bridging the gap between advanced computational models and clinical applicability, ultimately facilitating greater clinician confidence and usability in real-world settings [22].

The SHAP-based feature importance rankings observed in this study are broadly consistent with known pathophysiological mechanisms underlying LUTD subtypes. The dominance of bladder capacity in predicting OAB aligns with the established clinical understanding that reduced functional bladder capacity is a hallmark feature of this condition [2]. Similarly, the prominence of pelvic floor activity in predicting DV is coherent with the ICCS definition of dysfunctional voiding, which specifically requires evidence of habitual external urethral sphincter overactivity during the voiding phase [2, 26]. The finding that residual volume ranked as the second most influential predictor for DV further supports this interpretation, as elevated post-void residual is a recognized consequence of pelvic floor dyscoordination and incomplete bladder emptying.

A notable finding is the limited contribution of sex across all diagnostic categories in both classification settings. Although epidemiological studies have reported sex-based differences in LUTS prevalence (with functional incontinence and recurrent urinary tract infections more common in girls, and enuresis more evenly distributed [27, 28]) and normative uroflowmetry studies have demonstrated that Qmax and Qavg values are significantly higher in girls than in boys across all age groups [29], sex demonstrated relatively low SHAP values compared with the major predictive features in our models. This may indicate that sex-related differences in UF parameters are already captured indirectly through other features such as voided volume and Qmax, potentially rendering sex redundant in the presence of these covariates. Alternatively, because the diagnostic labels in our dataset

were assigned without sex-specific normative thresholds, sex may have been effectively neutralized during the labeling process itself. Future studies should investigate whether sex-stratified models or sex-specific normative cutoffs improve classification performance, particularly in younger age groups where anatomical and maturational differences may be more pronounced.

However, these findings must be interpreted with an important methodological caveat. Among the eight predictor features used in the models, three were derived from expert clinical interpretation rather than from objective device-generated measurements. Because the target diagnostic labels were also assigned by the same clinical experts based on the same uroflowmetry sessions, the high SHAP values observed for bladder capacity and pelvic floor activity may partly reflect this shared source of subjectivity rather than truly independent predictive information. This circularity does not invalidate the clinical coherence of the SHAP results, but it does limit the extent to which these findings can be interpreted as evidence of autonomous predictive capability. Future studies should evaluate model performance using only objectively measured, device-generated parameters to determine whether comparable classification accuracy can be achieved without clinician-interpreted inputs, thereby disentangling the model's independent contribution from the embedded clinical judgment.

Additionally, the inclusion of younger children in our cohort provides novel insights into UF characteristics in early childhood, a period that remains underrepresented in the literature. Early identification of lower urinary tract dysfunction has been linked to improved long-term renal and bladder outcomes [2]. By demonstrating that ML-based classification is feasible across a wide age range, including children under 5 years, our cohort's inclusion of children under 5 years highlights an underexplored population for future age-stratified model development and validation, although the present study did not perform a dedicated subgroup analysis to confirm model reliability in this age range.

Limitations

This study has a number of limitations that should be noted. First, among the eight predictor features used in the models, three were derived from expert clinical interpretation rather than from objective device-generated measurements. Because the target diagnostic labels were also assigned by the same clinical experts based on the same uroflowmetry sessions, the high SHAP values observed for bladder capacity and pelvic floor activity may partly reflect this shared source of subjectivity rather than truly independent predictive information. This circularity does not invalidate the clinical coherence of the SHAP results, but it does limit the extent to which these findings can be interpreted as evidence of autonomous predictive capability. Future studies should evaluate model performance using only objectively measured, device-generated parameters to determine whether comparable classification accuracy can be achieved without clinician-interpreted inputs, thereby disentangling the model's independent contribution from the embedded clinical judgment.

Second, this was a single-center retrospective study, and the generalizability of our findings to other institutions, populations, and UF devices is limited. Third, the DV-OAB class comprised only 68 records (2.55%), resulting in severe class imbalance that prevented several models from predicting this category entirely, as reflected by the undefined precision values reported in Table 4. This limitation reduces the clinical utility of the four-class setting and suggests that the DV-OAB subtype requires substantially larger sample sizes or alternative analytical strategies, such as oversampling, cost-sensitive learning, or one-versus-rest decomposition. Finally, ground-truth diagnostic labels were extracted from free-text clinical notes rather than from a prospectively standardized classification protocol. Although disagreements during initial labeling were resolved by consensus discussion between the two evaluating clinicians, a formal inter-rater agreement statistic (e.g., Cohen's kappa) was not calculated for the full dataset prior to consensus resolution. This approach may have introduced labeling inconsistencies, particularly for borderline cases. Future multicenter prospective studies with standardized labeling protocols, larger DV-OAB cohorts, and independent external validation are needed to confirm and extend these findings.

Conclusion

This study presented seven ML models for the classification of LUTD subtypes in a cohort of 2663 pediatric patients based on uroflowmetry data and associated clinical parameters. Taken together, our study expands the evidence base for UF in children, particularly those under 5 years of age. Our ML models demonstrate high consistency with clinical diagnoses; however, this alignment largely reflects the models' ability to reproduce the clinical criteria and expert interpretations used during the original diagnostic process. Therefore, rather than providing completely independent predictions, our explainable ML approaches demonstrate the potential to automate and standardize existing clinical decision-making. In light of these promising findings, future studies must evaluate models trained strictly on objective, device-generated variables to assess their capacity to truly reduce diagnostic subjectivity, conduct multicenter validation and use larger datasets to further confirm these findings

and to support the integration of such models into routine clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11255-026-05342-7.

Author contributions Conceptualization: N.D., S.M. and D.C.; Data curation: M.B., C.K.H., S.M., I.S., D.Ozmen and E.K.; Formal analysis: M.B., C.K.H., S.M., I.S., D.Ozmen and E.K.; Funding acquisition: D.Onur and N.D.; Investigation: N.D., S.M., D.Ozmen and E.K.; Methodology: D.Onur; Project administration: D.Onur and N.D.; Resources: D.Onur, N.D., S.M. and D.Ozmen; Software: D.Onur, M.B., C.K.H., S.M. and I.S.; Supervision: N.D., D.Ozmen and D.C.; Validation: D.Onur, M.B., C.K.H. and I.S.; Visualization: M.B., C.K.H. and I.S.; Writing—Original Draft: D.Onur, M.B., C.K.H., S.M., I.S. and E.K.; Writing—Review and Editing: N.D., S.M. and D.C.. All authors discussed the results and contributed to the final manuscript.

Funding This study was supported by the Health Institutes of Türkiye (TÜSEB), under Grant No. 2026-B1-01-52339.

Data availability All data supporting the findings of this study are available within the paper and its supplementary material. The datasets used in this study are not publicly available at the moment, but anonymised results are available on request.

Declarations

Conflict of interest The authors declare no potential conflicts of interest.

Ethical approval Approval was obtained from the University of Health Sciences İzmir Dr. Behçet Uz Children's Hospital Ethics Committee (279/2026/04–14). Informed consent was obtained from all participants and/or their legal guardians. All methods were conducted in accordance with the ethical standards set forth in the 1964 Declaration of Helsinki and its subsequent amendments.

Declaration of AI and AI-assisted technologies in the writing process In the course of composing the article, DeepL Translator & Write (DeepL SE, Cologne, Germany, www.deepl.com) and PoolText (PoolText Inc. Cambridge, United States, www.pooltext.com) were employed for the purposes of spelling, grammar, and reference accuracy. After using these tools/services, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bauer SB, Nijman RJM, Drzewiecki BA, Sillen U, Hoebeke P (2015) International Children's Continence Society standardization report on urodynamic studies of the lower urinary tract in children. Neurourol Urodyn 34:640–647. https://doi.org/10.1002/nau.22783
2. Austin PF, Bauer SB, Bower W, Chase J, Franco I, Hoebeke P et al (2016) The standardization of terminology of lower urinary tract function in children and adolescents: update report from the standardization committee of the International Children's Continence Society. Neurourol Urodyn 35:471–481. https://doi.org/10.1002/nau.22751
3. Van Batavia JP, Combs AJ (2018) The role of non-invasive testing in evaluation and diagnosis of pediatric lower urinary tract dysfunction. Curr Urol Rep 19:34. https://doi.org/10.1007/s11934-018-0784-1
4. Dudley AG, Casella DP, Lauderdale CJ, Zhao S, Chen H, Tanaka ST et al (2017) Interrater reliability in pediatric urodynamic tracings: a pilot study. J Urol 197:865–870. https://doi.org/10.1016/j.juro.2016.10.053
5. Vijverberg MAW, Klijn AJ, Rabenort A, Bransen J, Kok ET, Wingens JPM et al (2011) A comparative analysis of pediatric uroflowmetry curves. Neurourol Urodyn 30:1576–1579. https://doi.org/10.1002/nau.21152
6. Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 32:4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
7. Bentellis I, Guérin S, Khene Z-E, Khavari R, Peyronnet B (2021) Artificial intelligence in functional urology: how it may shape the future. Curr Opin Urol 31:385–390. https://doi.org/10.1097/MOU.0000000000000888
8. Ahmad I, Khondker A, Kwong JCC, Erdman L, Kim JK, Dos Santos J et al (2024) Artificial intelligence tools in pediatric urology: a comprehensive assessment of the landscape and current utilization. Curr Treat Options Pediatr 10:88–100. https://doi.org/10.1007/s40746-024-00301-9
9. Anjum MF, Karn A (2025) Small tools for small kidneys: artificial intelligence application in pediatric nephrology. Paediatr Nephrol J Bangladesh 10:75–79. https://doi.org/10.4103/pnjb.pnjb_12_25
10. Filler G, Gipson DS, Iyamuremye D, Díaz González De Ferris ME (2023) Artificial intelligence in pediatric nephrology—a call for action. Adv Kidney Dis Health 30:17–24. https://doi.org/10.1053/j.akdh.2022.11.001
11. Khondker A, Kwong JCc, Malik S, Erdman L, Keefe DT, Fernandez N et al (2022) The state of artificial intelligence in pediatric urology. Front Urol 2:1024662. https://doi.org/10.3389/fruro.2022.1024662
12. Khondker A, Ahmad I, Dhalla R, Kaushal S, Kwong JCC, Rickard M et al (2026) The 2025 update on artificial intelligence models in pediatric urology: Results from the AI-PEDURO collaborative. J Pediatr Urol. https://doi.org/10.1016/j.jpurol.2025.10.013
13. Nada A, Ahmed Y, Hu J, Weidemann D, Gorman GH, Lecea EG et al (2025) AI-powered insights in pediatric nephrology: current applications and future opportunities. Pediatr Nephrol. https://doi.org/10.1007/s00467-025-06911-1
14. Yücel ÖB, Tekin A, Tiryaki S, Mutlu O, Mert A, Ulman İ (2025) Artificial intelligence for objective assessment of pediatric uroflowmetry curves. Urology 206:125–130. https://doi.org/10.1016/j.urology.2025.08.022
15. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. https://doi.org/10.1136/bmj-2023-078378
16. Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern

Recognit Lett 30:27–38. https://doi.org/10.1016/j.patrec.2008.08.010
17. Lundberg S, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874. https://doi.org/10.48550/ARXIV.1705.07874
18. Van Batavia JP, Pohl HG, Farhat WA, Chiang G, BaniHani A, Collett-Gardere T et al (2023) Is it time to reconsider how we document pediatric uroflow studies?: a study from the SPU voiding dysfunction task force. J Pediatr Urol 19:546–554. https://doi.org/10.1016/j.jpurol.2023.04.018
19. Chang S-J, Yang SSD (2008) Inter-observer and intra-observer agreement on interpretation of uroflowmetry curves of kindergarten children. J Pediatr Urol 4:422–427. https://doi.org/10.1016/j.jpurol.2008.05.002
20. Sekerci CA, Tanidir Y, Tarcan T, Yucel S (2023) Consistency of uroflowmetry analysis in children among observers. Neurourol Urodyn 42:662–668. https://doi.org/10.1002/nau.25141
21. Arslan F, Algorabi O, Ozkan OC, Turkan YS, Namli E, Genc YE et al (2025) Artificial intelligence-based analysis of uroflowmetry patterns in children: a machine learning perspective. Neurourol Urodyn 44:1575–1582. https://doi.org/10.1002/nau.70139
22. Bang S, Tukhtaev S, Ko KJ, Han DH, Baek M, Jeon HG et al (2022) Feasibility of a deep learning-based diagnostic platform to evaluate lower urinary tract disorders in men using simple uroflowmetry. Investig Clin Urol 63:301. https://doi.org/10.4111/icu.20210434
23. Choo MS, Ryu HY, Lee S (2022) Development of an automatic interpretation algorithm for uroflowmetry results: application of artificial intelligence. Int Neurourol J 26:69–77. https://doi.org/10.5213/inj.2244052.026
24. Song SH, Choo M, Ryu HY, Lee S (2020) PD21-08 interpretation of uroflowmetrogram: application with artificial intelligence. J Urol 203:e456. https://doi.org/10.1097/JU.0000000000000871.08
25. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
26. Chase J, Austin P, Hoebeke P, McKenna P (2010) The management of dysfunctional voiding in children: a report from the Standardisation Committee of the International Children's Continence Society. J Urol 183:1296–1302. https://doi.org/10.1016/j.juro.2009.12.059
27. Nevéus T, Sillén U (2013) Lower urinary tract function in childhood; normal development and common functional disturbances. Acta Physiol 207:85–92. https://doi.org/10.1111/apha.12015
28. Salö M, Nejstgaard MC, Hambraeus M, Graneli C, Börjesson A, Hagelsteen K et al (2021) Sex and age differences in lower urinary tract dysfunction in healthy children. Acta Paediatr 110:2618–2626. https://doi.org/10.1111/apa.15959
29. Gupta DK, Sankhwar SN, Goel A (2013) Uroflowmetry nomograms for healthy children 5 to 15 years old. J Urol 190:1008–1014. https://doi.org/10.1016/j.juro.2013.03.073

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.