INTERNATIONAL JOURNAL OF
MEDICAL BIOCHEMISTRY

DOI: 10.14744/ijmb.2025.58815
Int J Med Biochem 2025;8(3):222–232

Research Article

Machine learning-assisted prediction of positive urine cultures using urinalysis and hemogram data: A retrospective cohort study

Ferhat Demirci[1], Yusuf Arikan[2], Ilkay Akbulut[3], Deniz Ilhan Topcu[4]

[1]Department of Medical Biochemistry, University of Health Sciences, Tepecik Training and Research Hospital, Izmir, Türkiye
[2]Department of Urology, University of Health Sciences, Tepecik Training and Research Hospital, Izmir, Türkiye
[3]Department of Infectious Diseases and Clinical Microbiology, University of Health Sciences, Tepecik Training and Research Hospital, Izmir, Türkiye
[4]Department of Medical Biochemistry, University of Health Sciences, Izmir City Hospital, Izmir, Türkiye

Abstract

Objectives: Urinary tract infections (UTIs) are common and often lead to unnecessary urine culture testing, increasing costs and delaying treatment. This study aims to develop a machine learning (ML) model using urinalysis and hemogram data to predict urine culture positivity and reduce unnecessary testing.

Methods: We retrospectively analyzed data from 12,433 patients who underwent urinalysis, urine culture, complete blood count, and CRP testing. After preprocessing and exclusion criteria, data were split into training, test, and validation sets. H_2O AutoML was employed to develop and evaluate various ML algorithms.

Results: The gradient boosting model demonstrated an AUC-ROC of 0.822 with high sensitivity (73.8%) and negative predictive value (90.4%), making it reliable in ruling out negative cases. Urinary leukocytes, nitrite, and bacterial count were identified as top predictors.

Conclusion: ML-based models can improve diagnostic accuracy and reduce unnecessary urine cultures. These models have the potential to be integrated into clinical workflows to enhance cost-effectiveness and minimize empirical antibiotic use.

Keywords: Artificial intelligence, diagnostic model, machine learning, predictive analytics, urinalysis, urinary tract infection, urine culture

How to cite this article: Demirci F, Arikan Y, Akbulut I, Topcu DI. Machine learning-assisted prediction of positive urine cultures using urinalysis and hemogram data: A retrospective cohort study. Int J Med Biochem 2025;8(3):222–232.

Urinary tract infections (UTIs) are among the most prevalent bacterial infections encountered in clinical practice, particularly affecting women. They impose a significant burden on both individual health and healthcare systems worldwide. If left untreated or inadequately managed, UTIs can lead to serious complications, including kidney damage and sepsis, while also contributing to the rise in antimicrobial resistance (AMR) rates [1, 2].

The diagnosis of UTIs involves a combination of clinical evaluation, laboratory tests, and advanced diagnostic techniques. Patients commonly present with symptoms such as dysuria, increased urinary frequency, urgency, and, in some cases, hematuria. Initial assessment typically includes microscopic urinalysis and dipstick testing, which together constitute standard urinalysis (UA). Although widely used, microscopic urinalysis is time-consuming and prone to human error. When combined

Address for correspondence: Ferhat Demirci, MD. Department of Medical Biochemistry, University of Health Sciences, Tepecik Training and Research Hospital, Izmir, Türkiye
Phone: +90 541 571 61 26 **E-mail:** drdemirci05@gmail.com **ORCID:** 0000-0002-5999-3399
Submitted: April 12, 2025 **Accepted:** June 01, 2025 **Available Online:** June 17, 2025

with dipstick tests—specifically leucocyte esterase and nitrite assessments—diagnostic accuracy improves significantly [3, 4].

Despite the utility of these methods, urine culture remains the gold standard for UTI diagnosis. However, it requires a minimum turnaround time of 24 hours, and culture positivity is observed in only 50–80% of suspected UTI cases. This variability results from multiple factors, including symptom severity and duration, patient characteristics, sample quality, etiological diversity, and methodological differences in testing [5–7]. Additionally, urine cultures are often unavailable in primary care settings and are typically limited to hospital microbiology laboratories. Consequently, empirical antibiotic therapy is frequently initiated based on clinical presentation and urinalysis results, with treatment adjustments made following culture outcomes [8].

The emergence of artificial intelligence (AI) holds promise for enhancing diagnostic accuracy and efficiency, potentially reducing delays in UTI diagnosis. While traditional diagnostic methods remain essential, the integration of AI could revolutionize the field by delivering more rapid and precise results. Importantly, AI-based diagnostics should complement—not replace—standard laboratory approaches to ensure comprehensive patient care [9]. These models leverage parameters such as patient age, bacterial presence, and specific analytical markers to accurately classify negative samples, thereby optimizing laboratory workflows and potentially reducing diagnostic costs by up to €40,000 annually [10].

This study aims to apply machine learning algorithms to urinalysis data in order to predict the necessity of urine culture testing. By doing so, we seek to minimize unnecessary culture requests and improve the accuracy of treatment decisions, ultimately contributing to more efficient and cost-effective management of UTIs.

Materials and Methods

Study population / subjects

This study was conducted at Tepecik Education and Research Hospital. Patients who presented to this center and its affiliated hospital (AH) between January 1, 2023, and December 31, 2023, and underwent first-time urinalysis (UA), urine culture, complete blood count (CBC), and C-reactive protein (CRP) testing were included. The baseline characteristics of the study population are presented in Table 1. Patients with incomplete test results, missing sub-parameters, or urine cultures identifying non-bacterial agents were excluded from the study.

Urinalysis samples were analyzed using the Zybio Corporation U2610 (Chongqing, China), CBC samples with the Sysmex Corporation XN-1000 (Kobe, Japan), and CRP testing with the Beckman Coulter AU-5800 (California, USA).

For urine culture, midstream samples were collected in sterile containers concurrently with urinalysis and processed following standard microbiological procedures. Samples with no growth signal after 24 hours were incubated for an additional 48 hours. If no growth was observed, the result was reported as "no growth." All reagents and calibrators used were certified and obtained from their respective manufacturers. Quality control materials were sourced from Bio-Rad (California, USA).

Study design

Ethical approval was obtained from the Tepecik Training and Research Hospital Ethics Committee prior to study initiation (No: 2024/07-13, Date: 19/08/2024). This study was performed in accordance with the ethical standards set by the Declaration of Helsinki. Patient identifiers were anonymized, and a dataset including age, sex, CRP, CBC, urinalysis, and urine culture results from 13,475 patients (12,085 from the main building and 1,390 from the affiliated hospital) was compiled using Microsoft Excel 2021 (USA).

After applying exclusion criteria, the final dataset included 12,433 patients (11,189 from the main hospital and 1,244 from the affiliated hospital).

In dipstick testing, semi-quantitative parameters were encoded as follows: 'negative'=0, 'trace'=0.5, and values 1, 2, or 3 for increasing levels of positivity. Urine color and appearance variables were also recategorized by merging similar classes to enhance data standardization.

The cleaned dataset was randomly divided into training and testing sets using an 80:20 ratio with stratified sampling to preserve class distribution. An additional external test set was used to evaluate model generalizability. The subject flow is outlined in the Standards for Reporting Diagnostic Accuracy (STARD) diagram (Fig. 1).

Data preprocessing and training of machine learning algorithms

Patient results were first exported to Microsoft Excel for initial preprocessing. Cases with missing values were excluded. Urine cultures with bacterial growth exceeding 10,000 colony-forming units/mL (CFU/mL) were classified as positive. Mixed flora, colonization, or growth below this threshold were labeled negative. Outcomes were binary coded: Negative (0), Positive (1).

Dipstick test results—such as glucose, protein, and nitrite—were converted into binary values. The final dataset was analyzed in Python 3.10 using the H_2O AutoML framework (version 3.46) [11]. AutoML was chosen due to its ability to automate complex processes such as feature engineering, model selection, and hyperparameter tuning—especially valuable when the user lacks deep data science expertise. Despite its growing relevance, AutoML has rarely been applied in clinical laboratory contexts [12].

Fifteen machine learning algorithms were evaluated (Appendix 1). The model with the highest AUC (area under the curve) was selected. The final model was trained on the following variables:

- Demographic variables: Age, sex,
- Hematologic variables: WBC (white blood cells), neutrophil, lymphocyte, monocyte, eosinophil, basophil, hemoglobin,

Table 1. Characteristics of the study population (no analysis was performed for semi-quantitative tests)

Characteristics	Units	Central Hospital (n=11189) Mean±SD	Affiliated Hospital (n=1244) Mean±SD	Reference interval
Age (years)		38.30±27.55	38.18±28.11	
Male (min-max)		40.11±28.86 (0–99)	38.74±29.26 (0–95)	
Female (min-max)		37.27±26.72 (1–104)	27.85±27.42 (1–95)	
Sex, n (%)				
Male		4065 (36.3)	462 (37.1)	
Female		7124 (63.7)	782 (62.9)	
WBC	$10^3/\mu L$	8.56±5.35	8.45±3.28	3.91–8.77
Neutrophil	$10^3/\mu L$	5.07±3.58	5.08±2.95	1.78–5.38
Lymphocyte	$10^3/\mu L$	2.56±3.68	2.46±1.34	0.85–3.0
Monocyte	$10^3/\mu L$	0.68±0.44	0.67±0.31	0.2–0.8
Eosinophil	$10^3/\mu L$	0.21±0.24	0.20±0.20	0.1–0.4
Basophil	$10^3/\mu L$	0.04±0.06	0.04±0.04	0.02–0.1
Hemoglobin	g/dL	12.23±1.89	12.19±1.86	11.9–15.4
Urine density	---	1018±0.007	1016±0.007	1010–1030
pH (urine)	---	5.96±0.77	5.99±0.77	5–9
Bacteria (urine)	/HPF	39.89±152.94	39.15±150.01	0–5
Leucocyte (urine)	/HPF	48.79±230.56	51.41±257.35	0–4
Yeast (urine)	/HPF	0.47±7.53	0.47±4.61	0
C-reactive protein	mg/L	10.42±40.57	8.40±33.88	0–5
Urine culture results	Train (+/-)	6700/2003 (77/23%)		
	Test (+/-)	1913/573 (77/23%)	958/286 (77/23%)	

SD: Standard deviation; min: Minimum, max: Maximum; WBC: White blood cells; HPF: High power field.

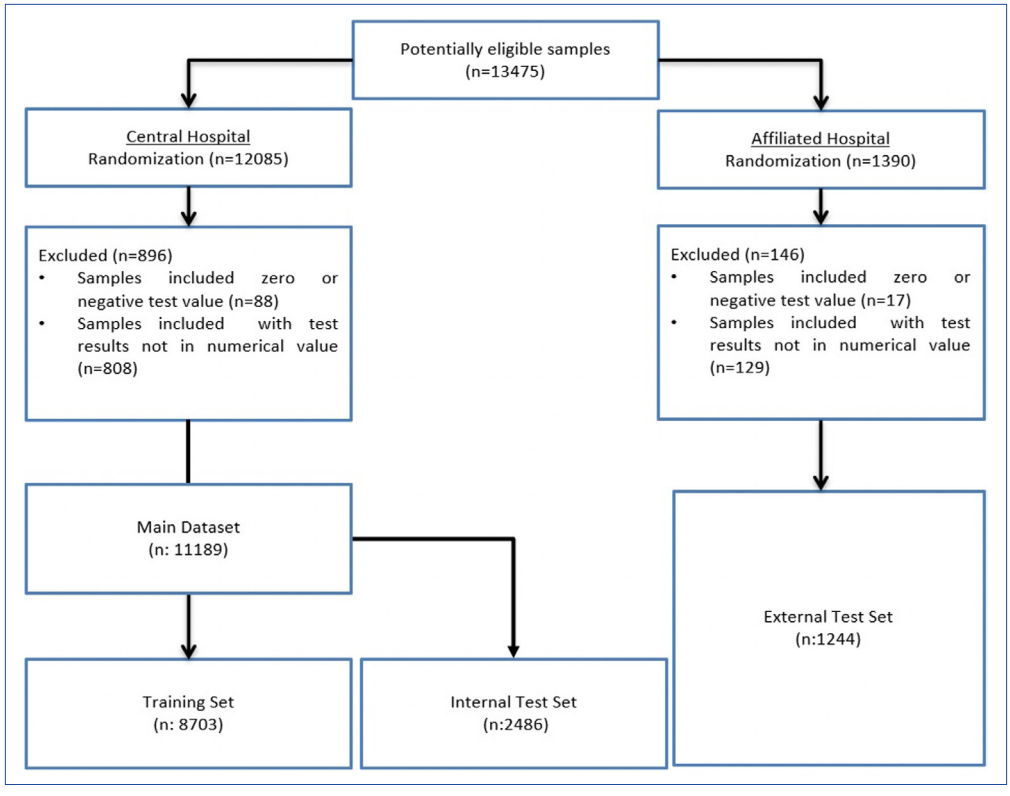

Figure 1. The standards for reporting diagnostic accuracy diagram.

- Urine dipstick variables: Appearance, urobilinogen, bilirubin, nitrite, ketone, leucocyte esterase, glucose, protein, pH, blood,
- Other urinalysis variables: Urine color, urine density, cylinder, mucus,
- Flow cytometry variable: Bacteria count, leucocyte count, yeast count.

Following model training, performance evaluation was conducted using the test dataset.

Performance evaluation

Scikit-learn, Pandas, NumPy, Shap, StatsModels, $H_2$0.automl and Matplotlib/Seaborn—among Python's most robust libraries for machine learning and statistical analysis—were employed in this project. The modeling process underwent comprehensive evaluation, including hyperparameter tuning and model selection through internal cross-validation. Model performance was assessed using multiple evaluation metrics. The following criteria were used for classification:

1. Classification performance metrics
    - Area Under the Receiver Operating Characteristic Curve (AUC-ROC),
    - Area Under the Precision-Recall Curve (AUC-PR),
    - Confusion matrix analysis,
    - Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV), Positive Likelihood Ratio (PLR), Negative Likelihood Ratio (NLR) F1 score, odds ratio.
2. Model interpretability metrics
    - Feature importance analysis,
    - SHAP (Shapley Additive Explanations) graphs.
3. Validation results of the predictive models were analyzed to ensure a comprehensive assessment. This structured and multifaceted evaluation approach provides a robust framework for predicting treatment modality outcomes based on laboratory-derived data. The algorithm was further validated using data from the affiliated hospital (external test set), which, while functioning as a distinct clinical site, operates under the same institutional umbrella as the central hospital. This approach aligns with the IFCC recommendations for assessing model generalizability across internal institutional subpopulations [13].

Results

Dataset description and data pre-processing

The dataset used in this study included a total of 11.189 records, consisting of 8.703 entries in the training set, 2.486 in the internal test set, and an additional 1.244 records in the external set. All datasets contained urinalysis and hemogram parameters alongside demographic data, allowing for comprehensive baseline characterization.

Baseline demographic characteristics of the study population are presented in Table 1. The mean age was 38.30±27.63 years in the training set, 38.33±27.28 years in the internal test set, and similar in the external test set at 38.18±28.11 years ($p>0.05$). When stratified by sex, male participants were slightly older than female participants within each subset ($p>0.05$).

Regarding sex distribution, males comprised 36.2% of the training set, 36.7% of the internal test set, and 37.1% of the external test set, while females made up 63.8%, 63.3%, and 62.9%, respectively. These differences were not statistically significant ($p=0.783$), indicating a relatively balanced gender distribution across the subsets.

Descriptive statistics for hemogram and urinary biomarkers are presented in Table 1. Most variables showed no statistically significant differences between the Central and Affiliated Hospital datasets—including blood WBC, neutrophil, monocyte, eosinophil, basophil, hemoglobin, CRP, urine density, pH, bacteria count, urinary leucocytes, and yeast (all $p>0.05$). Although a statistically marginal difference in lymphocyte counts was observed ($p=0.046$), the magnitude of difference was too small to be clinically meaningful. Overall, this observed homogeneity across subsets supports the robustness and comparability of subsequent analyses and model validation.

The performance of H_2O AutoML was comparatively evaluated based on predictive capabilities, classification metrics, and interpretability. Classification metrics such as F1 score, sensitivity, specificity, and AUC-ROC were used to assess the models' ability to discriminate between classes.

Comparison of classification performance metrics

The H_2O AutoML framework was employed to systematically explore a wide range of algorithms and hyperparameter configurations. Among the candidate models generated, a Gradient Boosting Machine (GBM) emerged as the most performant, striking an optimal trade-off between discrimination and calibration metrics—specifically AUC-ROC and log loss. The selected model (ID: GBM_1_AutoML_12_20250410_211225) achieved an AUC-ROC of 0. 0.818 and a log loss of 0.399, indicating both high classification accuracy and well-calibrated probabilistic outputs.

The performance metrics of the internal test set was summarized in Tables 2 and Figure 2a–c. The model achieved balanced classification performance with a sensitivity and specificity of 73.8%, and a high negative predictive value (NPV) of 90.4%, indicating strong reliability in ruling out negative cases. The positive predictive value (PPV) was 45.8%, and the resulting odds ratio of 7.95 further supported its overall discriminative capacity. Accuracy reached 73.8%, with an F1 score of 0.565, reflecting a moderate balance between precision and recall.

As visualized in Figure 2a, the model demonstrated an AUC-ROC of 0.822 and an AUC-PR of 0.649, confirming strong discriminative ability, particularly under class imbalance. The confusion matrix (Fig. 2c) further supports this consistent performance, underscoring the model's applicability in clinical diagnostic settings.

Table 2. Performance metrics of internal test set and external test set

Set	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Positive likelihood ratio	Negative likelihood ratio	Odds ratio	Accuracy	F1 score
Internal test set	0.738	0.738	0.458	0.904	2.819	0.355	7.948	0.738	0.565
External test set	0.738	0.735	0.453	0.904	2.783	0.357	7.798	0.736	0.562

External test set results of the models

In accordance with the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) recommendations, model validation was performed using data from the affiliated hospital to assess generalizability across institutional subpopulations. The results, summarized in Table 2 and Figure 2b, c, were consistent with the test sets, supporting the model's robustness and external applicability. Sensitivity and specificity were 73.8% and 73.5%, respectively, with a PPV of 45.3% and NPV of 90.4%, reflecting balanced

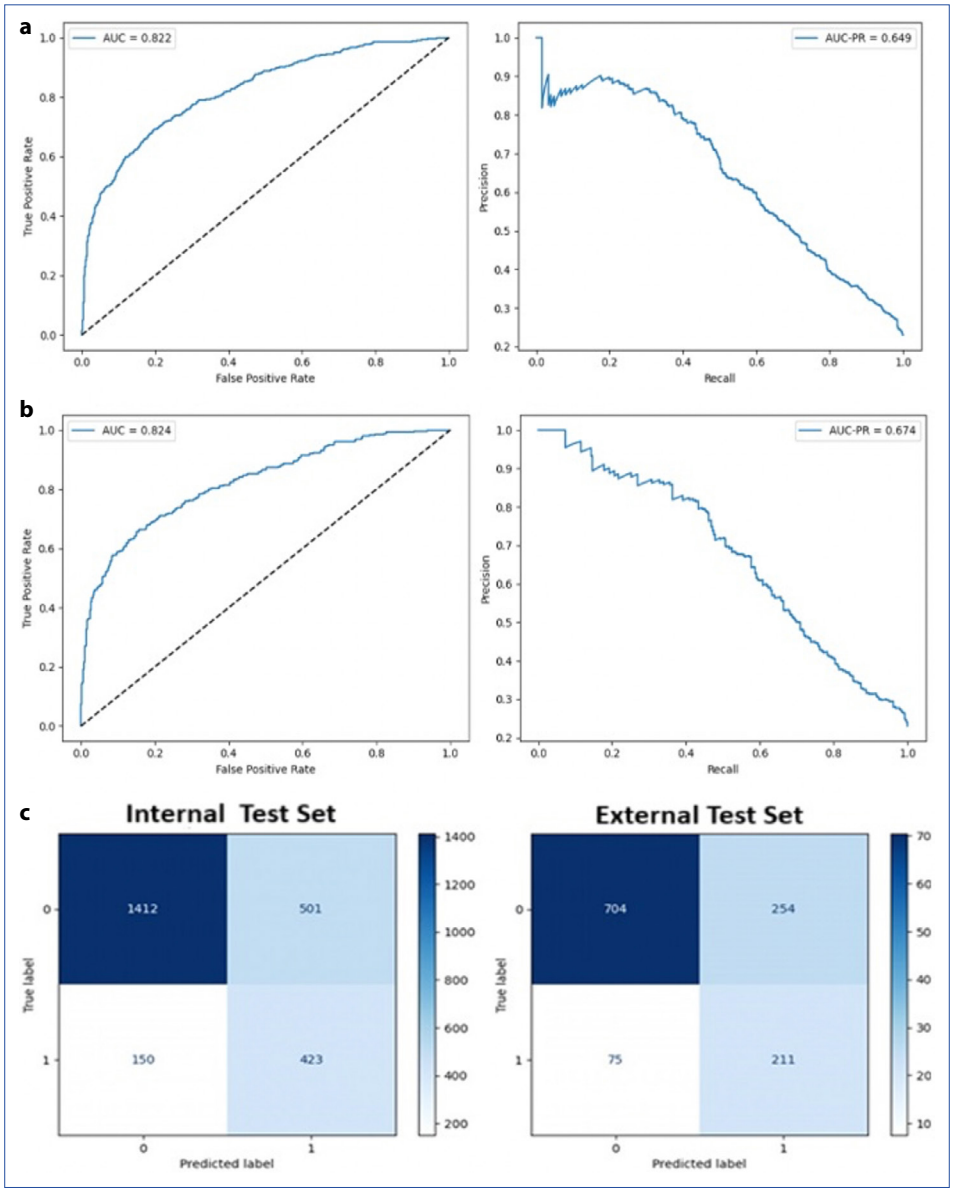

Figure 2. Performance outputs for our model. (a) AUC-ROC and AUC-PR graph for internal test set. (b) AUC-ROC and AUC-PR graph for external test set. (c) Confusion matrixes for internal and external test sets.

AUC: Area under the curve; ROC: Receiver operating characteristic; PR: Precision-recall.

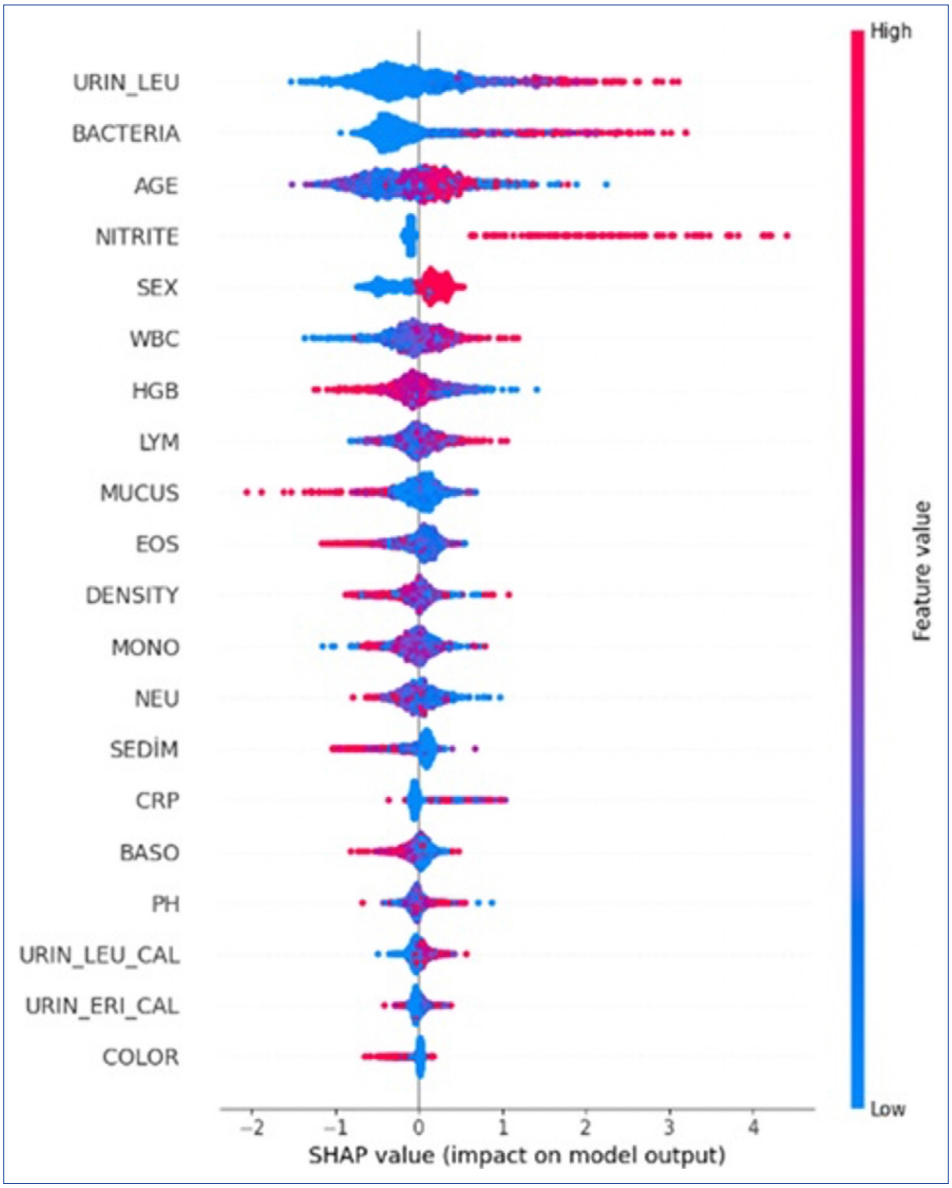

Figure 3. SHAP plot of our model.
SHAP: Shapley additive explanations.

classification and reliable identification of negative cases. The odds ratio (7.80), accuracy (73.6%), and F1 score (0.562) reinforced overall diagnostic utility.

As shown in Figure 2b, the model achieved an AUC-ROC of 0.824 and an AUC-PR of 0.674, slightly outperforming the test set and confirming its strong discriminative power. The confusion matrix (Fig. 2c) reflected similar prediction patterns, underscoring the model's consistency, generalizability, and clinical relevance within real-world healthcare settings.

Interpretability and threshold-based diagnostic performance of key variables

SHAP value analysis in Figure 3 highlights the most influential features driving the model's predictions. Urinary leucocyte (URIN_LEU) levels and bacterial count ranked highest, with elevated values strongly associated with increased likelihood of a positive classification—consistent with their well-established clinical importance in urinary tract infections. Age and nitrite also demonstrated substantial impact, particularly at higher levels. Moderate contributions were observed for sex, WBC, hemoglobin, and lymphocyte count, indicating context-specific influence on prediction.

Features such as urine pH, CRP, basophil, and sediment parameters had lower SHAP importance, though they may hold relevance in certain clinical subgroups. The distinct SHAP distributions reinforce the model's interpretability and alignment with biological plausibility.

Complementing SHAP analysis, Table 3 compares ROC-based and model-based diagnostic metrics for key infection-related variables commonly referenced in clinical

Table 3. Diagnostic metrics of key biomarkers

Variable	ROC cut-off	Sensitivity (ROC)	Specificity (ROC)	PPV (ROC)	NPV (ROC)	Sensitivity (model)	Specificity (model)	PPV (model)	NPV (model)
Leucocyte count (URIN_LEU)	12	0.561	0.906	0.781	0.776	0.563	0.819	0.482	0.863
Bacteria count	12	0.469	0.882	0.703	0.736	0.552	0.841	0.510	0.863
Leucocyte esterase (URIN_LEU_CAL)	1	0.688	0.804	0.677	0.812	0.668	0.706	0.404	0.877
Nitrite	1	0.198	0.997	0.979	0.676	0.266	0.981	0.809	0.817
Age	57	0.535	0.784	0.57	0.739	0.528	0.722	0.362	0.837

ROC: Receiver operating characteristic; PPV: Positive predictive value; NPV: Negative predictive value.

practice. Urinary leucocytes (URIN_LEU) demonstrated high ROC-based specificity (0.906) and PPV (0.781), alongside a strong model-based NPV (0.863), confirming its value in ruling out infection.

Leucocyte esterase (URIN_LEU_CAL), a point-of-care proxy for white blood cells, achieved robust standalone diagnostic performance (ROC sensitivity 0.688, specificity 0.804) and retained value in the model (NPV 0.877). Similarly, urinary bacteria showed high specificity (0.882) and consistent predictive power across both methods.

Although nitrite yielded an exceptionally high PPV (0.979) with ROC thresholds, its low sensitivity (0.198) and modest model contribution suggest it should be interpreted alongside complementary features. Age demonstrated moderate discriminative performance but provided consistent support across methods.

Together, these findings emphasize that routinely used clinical biomarkers—especially leucocyte esterase and urinary leucocytes—retain both individual and integrative predictive value within threshold-based and machine learning-driven diagnostic frameworks.

Discussion

This study evaluated the usability of machine learning models based on urinalysis in predicting the necessity of urine culture and classifying potential positive cases. The findings align with the literature supporting the clinical use potential of AI-based approaches in diagnosing urinary tract infections (UTIs).

Although symptomatology was not available in the dataset, the classification approach using $\geq 10^4$ CFU/mL as the threshold for positive urine cultures proved effective in the context of microbiological diagnostics. Notably, the model demonstrated strong performance in differentiating cases based on laboratory parameters alone. The high predictive contribution of urinary bacterial count suggests that the model could successfully identify patterns indicative of asymptomatic bacteriuria, catheter-associated infections, and uncomplicated cystitis. This outcome implies that even in the absence of clinical symptoms, machine learning algorithms can leverage routine urinalysis to support diagnostic differentiation across varied patient subgroups.

AI-based urinalysis risk scores have been shown to accelerate and improve diagnostic accuracy by predicting the need for culture testing, thereby reducing unnecessary diagnostics [14]. Integrating highly specific predictive models into clinical practice may help rationalize empirical antibiotic use and support more targeted treatment strategies [15]. By assessing key parameters such as leucocyte esterase, nitrite, and bacterial count, these systems offer significant cost savings for healthcare systems [16]. Future integration into clinical decision support systems is expected to enhance diagnostic workflows, improve patient safety, and contribute to infection control and the fight against antibiotic resistance. Recently, machine learning algorithms—particularly logistic regression, support vector machines, random forests, and deep learning—have shown promise in early UTI diagnosis by improving accuracy and reducing false positives, thus limiting unnecessary antibiotic use [3, 6, 7, 10]. The class imbalance observed in the dataset—characterized by a predominance of negative urine cultures—is a common feature in clinical laboratory data. To address this, stratified sampling was employed to ensure balanced class representation across all data splits. Furthermore, the use of H_2O AutoML provided automated internal handling of imbalance-related challenges and hyperparameter optimization, supporting reliable model calibration without the need for external resampling techniques. This workflow reflects real-world diagnostic settings and contributes to the model's practical applicability. These attributes collectively make H_2O AutoML an ideal framework for addressing the complexities of clinical laboratory data.

However, traditional studies have also reported significant findings on the diagnostic accuracy of urinalysis and microbiological tests. Price et al. [5] demonstrated that 30% of urine culture-positive patients were initially misdiagnosed as negative using dipstick tests. Similarly, Williams et all.'s [6] meta-analysis found that rapid urine tests had a sensitivity of 53–65% and a specificity of 85–90%.

In our study, CRP testing did not emerge as one of the most valuable features. It is well known that CRP levels do not typi-

cally rise in lower urinary tract infections due to the absence of a systemic inflammatory response [17]. Since patients in this study were not grouped based on symptoms or diagnosis, this phenomenon could not be explicitly explained.

Comparison of model performance with traditional methods

The best-performing model, GBM_1_AutoML_12_20250410_211225, demonstrated strong predictive performance on the test sets, with a sensitivity of 73.8%, specificity of 73.8%, positive predictive value (PPV) of 45.8%, negative predictive value (NPV) of 90.4%, positive likelihood ratio (PLR) of 2.82, negative likelihood ratio (NLR) of 0.36, an F1 score of 56.5%, and accuracy of 73.8%. The sensitivity of our model was lower compared to Heytens et al.'s [18] PCR-based analysis (70%).

In traditional studies, urinalysis-based diagnostic tests are often compared to urine cultures. Hooton et al. [19] reported that standard urinalysis in female patients had a sensitivity of 50–60% and specificity of 80–90%. Huysal et al. [20] found that routine laboratory tests had a sensitivity of 47% and specificity of 91.1%. Gupta et al. [21] indicated that dipstick tests were sufficient in terms of specificity for UTI diagnosis but had lower sensitivity (45–55%). These findings suggest that our model exhibits similar sensitivity to traditional methods but slightly lower performs them in specificity.

The reasons for these differences include the structural characteristics of the dataset used, variations in the patient population, and different preprocessing steps. Additionally, while traditional methods primarily employ univariate analyses, our model is based on multivariate analyses, which may result in higher sensitivity but lower specificity.

Comparison with other machine learning models

When compared with other machine learning models, Li et al.'s [22] machine learning-based models exhibited AUC-ROC values ranging from 0.68 to 0.97, sensitivity between 63–90%, and specificity between 69–86% [22]. The study by Burton et al. [23] in different patient groups reported an AUC-ROC value of 0.90, sensitivity ranging from 70–90.7%, specificity between 52–89%, accuracy of 63–85%, PPV of 40–71%, and NPV of 90–97%.

Seheult et al. [24] conducted studies using decision tree algorithms in different age groups, reporting an AUC-ROC range of 0.79–0.48, with an average sensitivity of 82.4%, specificity between 52–89%, accuracy of 65.8%, PPV of 46.3%, and NPV of 91.3%.

Flores et al. [25] developed a model combining neural networks and random forest algorithms, achieving an AUC-ROC of 0.81–85, sensitivity of 78–87%, specificity of 83%, accuracy of 80–85%, PPV of 86–83%, NPV of 74–87%, PLR of 4.6–5.07, and NLR of 0.26–0.16. Yen et al. [26] found the AUC-ROC value to be 0.83, sensitivity 88%, specificity 59%, accuracy 69.1%, F1 Score 65.2% in their study to identify high-risk patients with urinary tract infections that may cause critical outcomes in the emergency department. These studies had similar performance metric rates to our study.

While the model demonstrated high negative predictive value (NPV≈90%), which supports its utility as a rule-out tool, the relatively modest positive predictive value (PPV≈45%) indicates that a substantial number of predicted positive cases may not correspond to true infections. This imbalance raises important clinical considerations, particularly in the context of antibiotic stewardship and avoiding unnecessary interventions. However, we would like to emphasize that, from a patient safety perspective, missing a true infection (undertreatment) is clinically more critical than administering antibiotics unnecessarily. The potential financial and antimicrobial burden associated with overtreatment may be considered an acceptable trade-off when weighed against the risk of clinical deterioration due to untreated urinary tract infections.

Given the model's relatively modest positive predictive value (PPV≈45%) and moderate F1 score, its primary clinical utility may currently lie in ruling out infections rather than confirming them. The high negative predictive value (NPV≈90%) supports its role as a screening tool to exclude unnecessary culture testing in low-risk cases. Therefore, the model may be more effective as a "rule-out" aid in diagnostic workflows, helping reduce the burden of unwarranted laboratory procedures and antibiotic prescriptions until further improvements increase its confirmatory strength.

This trade-off between high NPV and lower PPV is common in diagnostic screening tools and reflects the real-world prevalence and distribution of urinary tract infections. Future improvements—such as threshold tuning, inclusion of symptom-based variables, or integration of additional inflammatory or microbiological markers—may enhance PPV without significantly compromising sensitivity, thereby expanding the model's practical applicability in clinical decision-making.

The model's interpretability, as assessed through SHAP analysis, revealed strong alignment with clinical intuition. Key variables such as urinary leucocytes, leucocyte esterase, and bacteria—which are already central to UTI diagnosis—emerged as the top contributors to prediction outcomes. Importantly, these features not only performed well in data-driven ML ranking but also retained their diagnostic strength under traditional ROC-based threshold analysis. This convergence highlights the potential of interpreting ML tools to bridge conventional clinical reasoning with algorithmic decision-making. The marginal performance of nitrite—despite high PPV but low sensitivity—further illustrates the value of multivariate modeling, where limitations of individual biomarkers can be mitigated by their collective interactions. These findings underscore the feasibility of using ML not just for black-box prediction, but as a transparent, clinically synergistic tool to enhance diagnostic efficiency in routine care.

Although urinalysis parameters are traditionally the primary focus in urinary tract infection diagnostics, this study also incorporated hemogram data to evaluate its additive predictive

value. Variables such as white blood cell count, lymphocyte percentage, and hemoglobin contributed moderately to the model's predictions according to SHAP analysis. While these parameters did not emerge as top-ranking features, their inclusion slightly improved the model's performance and may reflect systemic inflammatory responses in certain patient subgroups.

Notably, hemogram data are rarely emphasized in prior machine learning models for UTI detection. Our findings suggest that, although not dominant predictors, hematologic variables offer supplementary information that can enhance model robustness, particularly when urinalysis results are borderline or ambiguous. This reinforces the potential role of composite laboratory data in improving infection risk stratification through interpretable AI.

In studies conducted in recent years, AUC-ROC range, PPV/NPV, PLR/NLR balance were generally observed similar to our study. In contrast to this, although there are studies similar to the sensitivity/specificity balance in our study, there are studies that declare the opposite of these values. The main reasons for these differences include the variable selection of different machine learning models, hyperparameter optimization strategies, and the scope of the dataset used to train the model. Some studies included more clinical variables, while our study used only specific biomarkers. Furthermore, while some studies turn to deep learning methods, our model is based on traditional machine learning algorithms.

Limitations

This study has several limitations. First, as a retrospective analysis, the model's performance may vary across demographic subgroups and disease severity levels. Notably, females exhibited higher urine culture positivity rates than males across all subsets, as shown in Appendix 2, suggesting the need for future sex-stratified performance evaluations. Secondly, only urinalysis and basic hematologic data were used; clinical history, symptoms, and additional biomarkers were not included. Incorporating such features may improve model accuracy and applicability. Lastly, the integration of machine learning models into real-time clinical decision support systems and evaluation of their clinical impact remain essential future steps.

Conclusion

Our study shows that machine learning-based models can be effective in the early diagnosis of urinary tract infections. Although the sensitivity of our model is lower compared to some studies, its specificity is quite high. This suggests that the model can prevent unnecessary antibiotic use by reducing false positives. Future studies should test the model in different patient groups, add symptomatic data and validate it in real-time clinical applications. The recommendations of the IFCC working group on the application of artificial intelligence in laboratory medicine also suggest cautious application of these technologies in a clinical context. In this context, additional studies are needed to improve the usability of machine learning models in the hospital setting.

Ethics Committee Approval: The study was approved by the Tepecik Training and Research Hospital Non-interventional Ethics Committee (no: 2024/07-13, date: 19/08/2024).

Informed Consent: Informed consent was obtained from all participants.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Funding: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: No AI technologies utilized.

Authorship Contributions: Concept – F.D., D.I.T.; Design – F.D., D.I.T.; Supervision – F.D., I.A., Y.A.; Materials – F.D., I.A., Y.A.; Data collection and/or processing – F.D., I.A., Y.A.; Data analysis and/or interpretation – F.D., D.I.T.; Literature search – F.D., D.I.T.; Writing – F.D., D.I.T.; Critical review – F.D., D.I.T.

Peer-review: Externally peer-reviewed.

References

1. Sarkar R, Wondwosen B, Fikru A, Shume T. An overview of urinary tract infection. In AK Prajapati editor. Microbes of Medical Importance. Iterative International Publishers; 2024. p. 641–78. [CrossRef]
2. Bavanandan S, Keita N. Urinary tract infection prevention and treatment. Semin Nephrol 2023;43(5):151468. [CrossRef]
3. Hans A, Yadav A, Kaur P, Kumari A. Evaluation of leukocyte esterase and nitrite dipstick tests with routine urine microscopic analysis in detecting urinary tract infections. Indian J Pathol Oncol 2024;11(1):3–7. [CrossRef]
4. Liou N, De T, Urbanski A, Chieng C, Kong Q, David AL, et al. A clinical microscopy dataset to develop a deep learning diagnostic test for urinary tract infection. Sci Data 2024;11(1):155. [CrossRef]
5. Price TK, Dune T, Hilt EE, Thomas-White KJ, Kliethermes S, Brincat C, et al. the clinical urine culture: Enhanced techniques improve detection of clinically relevant microorganisms. J Clin Microbiol 2016;54(5):1216–22. [CrossRef]
6. Williams GJ, Macaskill P, Chan SF, Turner RM, Hodson E, Craig JC. Absolute and relative accuracy of rapid urine tests for urinary tract infection in children: A meta-analysis. Lancet Infect Dis 2010;10(4):240–50. [CrossRef]
7. Heytens S, De Sutter A, Coorevits L, Cools P, Boelens J, Van Simaey L, et al. Women with symptoms of a urinary tract infection but a negative urine culture: PCR-based quantification of Escherichia coli suggests infection in most cases. Clin Microbiol Infec 2017;23(9):647–52. [CrossRef]
8. Luciano R, Piga S, Federico L, Argentieri M, Fina F, Cuttini M, et al. Development of a score based on urinalysis to improve the management of urinary tract infection in children. Clinica Chimica Acta 2012;413(3–4):478–82. [CrossRef]
9. Mshana NG, Choudhary S, Garg VK. Artificial intelligence and urinary tract infections: A diagnostic perspective. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2024. p. 1–6. [CrossRef]
10. Del Ben F, Da Col G, Cobârzan D, Turetta M, Rubin D, Buttazzi P, et al. A fully interpretable machine learning model for in-

creasing the effectiveness of urine screening. Am J Clin Pathol 2023;160(6):620–32. [CrossRef]
11. Fryda T, LeDell E, Gill N, Aiello S. H2O: R Interface for the "H2O" scalable machine learning platform. Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/index.html. Accessed Feb 10, 2025.
12. Topcu Dİ, Bayraktar N. Searching for the urine osmolality surrogate: An automated machine learning approach. Clin Chem Lab Med 2022;60(12):1911–20. [CrossRef]
13. Master SR, Badrick TC, Bietenbeck A, Haymond S. Machine learning in laboratory medicine: Recommendations of the IFCC Working Group. Clin Chem 2023;69(7):690–8. [CrossRef]
14. Choi MH, Kim D, Bae HG, Kim AR, Lee M, Lee K, et al. Predictive performance of urinalysis for urine culture results according to causative microorganisms: An integrated analysis with artificial intelligence. J Clin Microbiol 2024;62(10):e0117524. [CrossRef]
15. Impana KP, Saya A, Shetty BH, Manaswini C, Ashjay CA. Survey on machine learning models to analyze urinary tract infection data. Int Res J Adv Eng Manag 2024;2(04):1097–109. [CrossRef]
16. Farashi S, Momtaz HE. Prediction of urinary tract infection using machine learning methods: A study for finding the most-informative variables. BMC Med Inform Decis Mak 2025;25(1):13. [CrossRef]
17. Narayan Swamy SN, Jakanur RK, Sangeetha SR. Significance of C-reactive protein levels in categorizing upper and lower urinary tract infection in adult patients. Cureus 2022;14(6):e26432. [CrossRef]
18. Heytens S, De Sutter A, Coorevits L, Cools P, Boelens J, Van Simaey L, et al. Women with symptoms of a urinary tract infection but a negative urine culture: PCR-based quantification of Escherichia coli suggests infection in most cases. Clin Microbiol Infec 2017;23(9):647–52. [CrossRef]
19. Hooton TM, Roberts PL, Cox ME, Stapleton AE. Voided midstream urine culture and acute cystitis in premenopausal women. New England J Med 2013;369(20):1883–91. [CrossRef]
20. Huysal K, Budak YU, Ulusoy Karaca A, Aydos M, Kahvecioğlu S, Bulut M, et al. Diagnostic accuracy of UriSed automated urine microscopic sediment analyzer and dipstick parameters in predicting urine culture test results. Biochem Med (Zagreb) 2013;23(2)211–7. [CrossRef]
21. Gupta K, Trautner BW. The 2019 USPSTF report on screening for asymptomatic bacteriuria-lessons from history. JAMA Netw Open 2019;2(9):e1912522. [CrossRef]
22. Li J, Du Y, Huang G, Zhang C, Ye Z, Zhong J, et al. Predictive value of machine learning model based on CT values for urinary tract infection stones. iScience 2024;27(12):110843. [CrossRef]
23. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. BMC Med Inform Decis Mak 2019;19(1):171. [CrossRef]
24. Seheult JN, Stram MN, Contis L, Pontzer RE, Hardy S, Wertz W, et al. Development, evaluation, and multisite deployment of a machine learning decision tree algorithm to optimize urinalysis parameters for predicting urine culture positivity. J Clin Microbiol 2023;61(6):e0029123. [CrossRef]
25. Flores E, Martínez-Racaj L, Blasco Á, Diaz E, Esteban P, López-Garrigós M, et al. A step forward in the diagnosis of urinary tract infections: From machine learning to clinical practice. Comput Struct Biotechnol J 2024;24:533–41. [CrossRef]
26. Yen CC, Ma CY, Tsai YC. Interpretable machine learning models for predicting critical outcomes in patients with suspected urinary tract infection with positive urine culture. Diagnostics. 2024;14(17):1974. [CrossRef]

Appendix 1

	Model_id	AUC	Logloss
1	GBM_1_AutoML_12_20250410_211225	0.818256	0.398633
2	GBM_5_AutoML_12_20250410_211225	0.816776	0.400571
3	GBM_2_AutoML_12_20250410_211225	0.816195	0.402202
4	XGBoost_grid_1_AutoML_12_20250410_211225_model_1	0.815258	0.400285
5	GBM_grid_1_AutoML_12_20250410_211225_model_1	0.811711	0.402868
6	GBM_3_AutoML_12_20250410_211225	0.809752	0.407616
7	DRF_1_AutoML_12_20250410_211225	0.809508	0.442237
8	XGBoost_3_AutoML_12_20250410_211225	0.808496	0.471458
9	XGBoost_grid_1_AutoML_12_20250410_211225_model_2	0.807497	0.409993
10	GBM_4_AutoML_12_20250410_211225	0.807266	0.415546
11	XGBoost_grid_1_AutoML_12_20250410_211225_model_3	0.805151	0.414008
12	GBM_grid_1_AutoML_12_20250410_211225_model_2	0.804209	0.419474
13	XRT_1_AutoML_12_20250410_211225	0.800781	0.445000
14	XGBoost_1_AutoML_12_20250410_211225	0.794679	0.475173
15	DeepLearning_grid_1_AutoML_12_20250410_211225_...	0.794134	0.434181

H_2O AutoML models and performance metrics. AUC: Area under the curve.

Appendix 2

Dataset-sex	Negative (n)	Positive (n)	Negative (%)	Positive (%)
Train set-male	2571.0	582.0	81.54	18.46
Train set-female	4129.0	1421.0	74.40	25.60
Internal test set-male	736.0	176.0	80.70	19.30
Internal test set-female	1177.0	397.0	74.78	25.22
External test set-male	374.0	88.0	80.95	19.05
Externa test set-female	584.0	198.0	74.68	25.32

Urine culture distribution by sex and dataset.