Ferhat Demirci https://orcid.org/0000-0002-5999-3399
Murat Akşit https://orcid.org/0000-0003-2106-9130
Ilkay Akbulut https://orcid.org/0000-0002-4840-6865
Aylin Demirci https://orcid.org/0000-0002-4759-1990

RESEARCH ARTICLE / ARAŞTIRMA
DOI: 10.4274/mjima.galenos.2025.25555.4
Mediterr J Infect Microb Antimicrob 2026;15:25555.4
Erişim: http://dx.doi.org/10.4274/mjima.galenos.2025.25555.4

A Laboratory Decision-Support System for Reflective Urine Culture Testing: Development of an Interpretable AI Model

Yansıtmalı İdrar Kültürü Testi için Laboratuvar Karar Destek Sistemi: Yorumlanabilir bir Yapay Zekâ Modeli Geliştirilmesi

Demirci et al. Laboratory Decision-Support System for Reflective Urine Testing

Ferhat Demirci[1,2], Murat Akşit[2], İlkay Akbulut[3], Aylin Demirci[4]

[1]Dokuz Eylül University, Institute of Health Sciences, Department of Neurosciences, İzmir, Türkiye
[2]University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Medical Biochemistry, İzmir, Türkiye
[3]University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Infectious Diseases and Clinical Microbiology, İzmir, Türkiye
[4]University of Health Sciences Türkiye, İzmir Tepecik Training and Research Hospital, Department of Family Medicine, İzmir, Türkiye

Ferhat Demirci MD, Dokuz Eylül University, Institute of Health Sciences, Department of Neurosciences, İzmir, Türkiye
drferhat5505@hotmail.com
0000-0002-5999-3399

Cite this article as: Demirci F, Akşit M, Akbulut İ, Demirci A. A laboratory decision-support system for reflective urine culture testing: Development of an interpretable AI model. Mediterr J Infect Microb Antimicrob.

18.07.2025
08.12.2025

Epub: 07.01.2026
Published:

Abstract
Introduction: Urinary tract infections are a common diagnostic challenge. Although urine culture remains the gold standard, it is time-consuming and often ordered reflexively. This study aimed to develop and validate an interpretable machine-learning–based Laboratory Decision-Support System (LDSS) to guide reflective urine culture prioritization using only structured laboratory data.
Materials and Methods: We analyzed a retrospective cohort of 51,923 adult patients. Seven machine learning algorithms were trained, with the random forest (RF) model demonstrating the highest accuracy. SHapley Additive exPlanations analysis was employed to ensure model interpretability. A reduced RF model, using the top 10 predictive features, was used to construct three scoring systems: one emphasizing model fidelity, one optimizing diagnostic balance, and one maximizing sensitivity.
Results: The RF model demonstrated excellent performance (external receiver operating characteristic–area under the curve [ROC-AUC]: 0.956). The simplified 10-variable model maintained high accuracy (ROC–AUC: 0.947). Key predictors included bacterial count, leukocyte count, nitrite presence, and patient age. The scoring systems offered flexible options tailored to different diagnostic priorities, with the SAFE-Score achieving 95.3% sensitivity.
Conclusion: The LDSS is intended to support reflex culture prioritization, not reduce overall culture testing. By streamlining pre-analytical triage and highlighting clinically significant samples, it promotes appropriate culture utilization and strengthens antimicrobial stewardship, while preserving the central role of urine culture in infection management.
Keywords:

Introduction

Urinary tract infections (UTIs) are among the most common infections in clinical practice, with an estimated global incidence exceeding 150 million cases annually [1]. They are associated with substantial healthcare costs, frequent antibiotic prescriptions, and increased diagnostic burden, particularly in outpatient and emergency settings [2,3]. Accurate diagnosis remains challenging due to nonspecific symptoms and reliance on time-consuming laboratory tests [4].

Urine culture is considered the gold standard for UTI diagnosis. However, its 24–48-hour turnaround often necessitates empiric antibiotic treatment before microbiological confirmation [5]. This practice contributes to antimicrobial resistance, now recognized by the World Health Organization as a global health threat [6]. Moreover, up to 60%–70% of urine cultures yield negative or clinically insignificant results, highlighting potential overuse of testing and therapy [7].

Rapid dipstick tests, detecting leukocyte esterase and nitrite, provide immediate screening but show variable performance across populations, with sensitivity and specificity ranging from 68% to 88% and 17% to 98%, respectively [8].
This diagnostic uncertainty has prompted efforts to improve laboratory decision-making, including the use of reflective testing. Reflective testing, increasingly recognized in modern laboratory medicine, involves laboratory physicians adding further analyses or interpretative comments after reviewing initial test results to enhance diagnostic reasoning [9]. In UTIs, this expert-led approach aids accurate interpretation and encourages more judicious use of microbiological testing. Laboratory physicians thus face the dual challenge of minimizing unnecessary culture requests while ensuring patients with a high likelihood of positive cultures are correctly identified.

In most laboratory information systems (LIS), detailed symptom information is not captured; only test orders and preliminary diagnoses, such as International Classification of Diseases (ICD) codes, are typically available. Consequently, the predictive modeling approach in this study relied solely on structured laboratory data. To address this, we developed a standardized, interpretable, and data-driven Laboratory Decision-Support System (LDSS) to optimize urine culture utilization using routine laboratory parameters. The LDSS is not intended to replace clinical diagnoses but to assist laboratory physicians in prioritizing reflex urine culture testing within laboratory workflows. Diagnostic responsibility remains entirely with the treating clinician, while the LDSS provides reproducible, standardized insights derived from LIS data.

Artificial intelligence (AI) and machine learning (ML) have gained increasing attention for developing predictive models in UTI diagnosis. Various algorithms—including logistic regression, random forests (RFs), XGBoost, Light Gradient Boosting Machine (LightGBM), and TabNet—have demonstrated robust performance using structured data such as urinalysis results, demographics, and clinical history [10–12]. Reported AUROC values commonly exceed 0.85, with some studies achieving 0.95 or higher in external validation cohorts [11,13].

Recent studies have highlighted the importance of model interpretability. By employing SHapley Additive exPlanations (SHAP), our LDSS not only ensures transparency but also facilitates clinical integration by illustrating the real-time contribution of each variable. Real-world implementations of ML-based LDSSs have shown reductions in unnecessary culture orders, accelerated treatment decisions, and improved antibiotic stewardship outcomes [12,14].

Despite these advances, challenges remain. Many predictive models are trained on single-center datasets and lack external validation, raising concerns about generalizability across institutions and diverse patient populations [13,15]. Additionally, variability in urinalysis platforms and clinical practice patterns may limit reproducibility and scalability. Unlike existing tools, the proposed LDSS provides three distinct scoring systems tailored to different clinical priorities, ranging from high-sensitivity triage to specificity-focused decision-making. This flexibility promotes collaboration among biochemists, microbiologists, and clinicians while reducing diagnostic waste by minimizing unnecessary urine culture requests.

The aim of this study was to develop and externally validate a robust, interpretable ML-based LDSS to predict urine culture outcomes in patients with suspected UTIs. By standardizing reflective testing practices, the LDSS supports interdisciplinary decision-making, optimizes resource utilization, and ultimately contributes to rational antibiotic prescribing across healthcare settings.

Materials and Methods

Study Population/Subjects

This study was conducted at xxxx Hospital. Ethical approval was obtained from the xxxx Hospital Ethics Committee prior to study initiation (Resolution No. 2025/02-05, dated March 10, 2025).

Eligible participants were adults aged ≥18 years who presented as inpatients or outpatients to the main hospital between January 1, 2014, and December 31, 2024, or to its affiliated hospital between January 1 and February 28, 2025. Inclusion criteria required patients to undergo their first urinalysis, complete blood count (CBC), and urine culture, ordered by a specialist physician based on clinical indication.

The study cohort included both culture-positive and culture-negative cases, capturing the full spectrum of patients for whom urine cultures were clinically indicated. Consequently, the dataset reflects real-world test-ordering practices rather than a biased subset of confirmed infections.

Patients were excluded if they had incomplete test results, missing sub-parameters, non-bacterial pathogens in their urine culture, delays exceeding one hour between urine sample collection and laboratory registration, delays exceeding 30 minutes for hemogram samples between phlebotomy and laboratory receipt, or a history of antibiotic treatment prior to testing.

CBC analyses were performed using UniCell DxH 800 analyzers (Beckman Coulter, Miami, FL, USA) from 2014 to 2020 and XN-2000 systems (Sysmex Corporation, Kobe, Japan) from 2020 onward. Urinalysis tests were conducted using fully automated analyzers across three periods: H-800 and FUS-200 systems (Dirui Industrial Co., Changchun, China) from 2014 to 2018; BT Uricell 1280–1600 (Bilimsel Products, Izmir, Türkiye) from 2018 to 2021; and U2610–U1600 (Zybio Corporation, Chongqing, China) from 2021 onward.

Midstream urine samples were collected in sterile containers simultaneously with urinalysis and processed according to standard microbiological procedures. Samples without detectable bacterial growth after 24 hours were incubated for an additional 48 hours; if no growth was observed, the result was reported as "no growth."

Reagents and calibrators for urinalysis were obtained from authorized manufacturers and were certified and registered products. Quality control materials were sourced from Bio-Rad (California, USA). All results were reviewed and validated for accuracy and reliability by both a clinical biochemistry specialist and a clinical microbiology specialist.

Study Design

Patient identifiers were anonymized, and a dataset comprising age, sex, hemogram, urinalysis, and urine culture results from 55,385 patients (main hospital: 52,854; affiliated hospital: 2,531) was imported into Microsoft Excel 2021 (USA). Symptom data were not included, as such information is not routinely recorded in LIS. In standard laboratory workflows, test orders are typically accompanied by preliminary diagnoses or ICD codes from the requesting physician, but detailed patient symptoms are not captured. Accordingly, the predictive model in this study was developed exclusively on structured laboratory data, aiming to forecast urine culture outcomes rather than to establish a clinical diagnosis of UTI. After applying exclusion criteria, the final dataset included 49,720 patients, with an external validation cohort of 2,203 patients. The dataset was subsequently transferred to Python (version 3.13.1, USA) for ML analysis.
Following data cleaning, the main dataset was divided into training, internal test, and external test subsets using a 60:20:20 stratified sampling strategy based on the binary target variable, ensuring preservation of class distribution. Patient flow throughout the study is depicted in Figure 1, in accordance with the Standards for Reporting Diagnostic Accuracy (STARD) guidelines.

Data Preprocessing and Training of ML Algorithms

Patient data were initially exported from the LIS into Microsoft Excel. Hemogram values and flow cytometry parameters from urinalysis were used directly due to device standardization. Semi-quantitative dipstick results—reported by urinalysis analyzers as categorical values (e.g., "+," "++," "+/-," "trace")—were converted into numerical equivalents (e.g., "++" mapped to 2; "trace" standardized to 0.5) to ensure quantitative consistency. Variables describing urine color and appearance were also recategorized by grouping similar classifications (e.g., light yellow to dark red; clear to very cloudy) to standardize the dataset.

Urine culture results were binarized as follows: samples with ≥10,000 CFU/mL bacterial growth were defined as positive (label = 1), while samples with <10,000 CFU/mL, mixed flora, colonization, yeast, or no growth were classified as negative (label = 0).

The 10,000 CFU/mL threshold was selected based on recent evidence and the 2024 European Association of Urology guidelines, which acknowledge that lower colony counts ($\geq 10^3$–10^4 CFU/mL) may be clinically significant in symptomatic or catheterized patients. Nelson et al. demonstrated that these lower thresholds preserve diagnostic accuracy for symptomatic UTIs, supporting their use in reflective testing workflows. Additionally, Werneburg et al. showed that urinalysis parameters reliably predict the absence of infection at this threshold, reinforcing its clinical validity. This definition also aligns with our institutional microbiology reporting standard for significant bacteriuria [16–18].

Yeast and colonization findings were labeled as negative (label = 0) based on established microbiological evidence and laboratory reporting standards. In urinary cultures, the presence of Candida species typically reflects colonization or contamination rather than true infection, even at colony counts exceeding 10^4–10^5 CFU/mL, unless accompanied by compatible clinical symptoms [19]. Classifying yeast as negative prevented false-positive propagation in the LDSS and improved the model's clinical specificity.

Similarly, cases labeled as "colonization"—including cultures with mixed flora or non-uropathogenic organisms—were considered negative. This approach aligns with standard microbiology practice, where such findings are reported as clinically non-significant. Although CLSI M100 (2025) does not define colony-count thresholds for colonization or candiduria, its terminology guided our categorization strategy. This interpretation reflects real-world laboratory workflows, ensuring that the LDSS mirrors standardized reporting logic and remains generalizable across institutions [20].

The cleaned dataset was transferred to Python for ML analysis. To enhance model robustness and address class imbalance, a stratified data partitioning scheme was applied, allocating 60% of samples to training and 20% each to internal and external testing. The dataset exhibited natural imbalance, with 22.4% culture-positive and 77.6% culture-negative samples. To mitigate majority-class bias, feature standardization and rebalancing strategies (class_weight='balanced') were applied uniformly across all classifiers.

As a preliminary check, a baseline Logistic Regression model was trained and evaluated across all data splits. ROC-AUC scores (≈0.74, 0.73, 0.73 for training, internal, and external sets, respectively) and F1 scores (0.55, 0.54, 0.54) demonstrated consistent generalization without evidence of overfitting or imbalance-driven inflation. The close alignment of these baseline metrics confirmed that stratified sampling preserved class proportions across all subsets (≈22.4% positive vs. 77.6% negative), ensuring reliable model development.

ML Model Selection and Development
The results confirmed that the methodological setup—including stratified sampling and proportional weighting—effectively mitigated class imbalance and provided a reliable foundation for model development. Logistic Regression was used not as a primary model, but as a diagnostic tool to verify dataset integrity and the fairness of the training process [21].

Model development was performed in Python 3.13.1 using widely adopted libraries and workflows. Seven ML algorithms were evaluated for their suitability with the dataset and their potential effectiveness in predicting urine culture outcomes: RF, Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, Logistic Regression (LR), Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN).

Variables included in the analysis:

- **Demographic:** Age, sex
- **Hemogram:** White blood cell, neutrophil, lymphocyte, monocyte, eosinophil, basophil, hemoglobin (HGB)
- **Urine Dipstick:** Leukocyte esterase, nitrite, glucose, protein, pH, erythrocyte, bilirubin, urobilinogen, ketone
- **Other Urinalysis:** Urine color, urine density, appearance
- **Flow Cytometry:** Bacteria count, cylinder, yeast, urine leukocyte count

Data preprocessing, model training, evaluation, and visualization were conducted using open-source Python libraries:

- **Data Processing and Analysis:** pandas (v2.2.2), numpy (v2.0.2), optuna (v4.3.0)
- **ML Model Development:** scikit-learn (v1.6.1), xgboost (v2.1.4), lightgbm (v4.5.0), catboost (v1.2.8), tensorflow (v2.10), keras (v2.10), torch (v2.6.0 + cu124)
- **Model Evaluation and Visualization:** matplotlib (v3.10), seaborn (v0.13.2), scipy.stats (v1.9), sklearn.metrics (v1.2), SHAP (v0.47)

Detailed hyperparameter optimization procedures, including search strategies and parameter configurations for each model, are provided in the Supplementary Material (Table S1). Each model was retrained using the optimal hyperparameters identified during tuning. Final model evaluation was based on F1 and ROC-AUC scores derived from the internal test set.

Performance Evaluation
Performance evaluation was conducted using standard Python-based data science libraries. The modeling process was assessed comprehensively through internal cross-validation, hyperparameter tuning, and multiple performance metrics.

1. Classification Performance Metrics: Model discrimination and predictive capability were evaluated using:

- Area Under the Receiver Operating Characteristic Curve (AUC–ROC)
- Area Under the Precision-Recall Curve (AUC-PR)
- Sensitivity and Specificity
- Positive Predictive Value (PPV) and Negative Predictive Value (NPV)
- Positive Likelihood Ratio (PLR) and Negative Likelihood Ratio (NLR)
- F1-Score

2. Model Interpretability Metrics: To enhance clinical transparency and foster trust in algorithmic decisions, interpretability was assessed using:

- Feature-Importance metrics
- SHAP graphs

This multidimensional evaluation approach balances predictive performance with explainability, providing a robust framework for forecasting urine culture outcomes based solely on laboratory and demographic data.

Development of the LDSS
The LDSS was built using the best-performing ML model identified during model selection. SHAP analysis was employed to select the ten most informative features, and a simplified model was retrained using only these variables. The reduced model maintained performance comparable to the full model, supporting its suitability for practical implementation. Instead of the default probability threshold of 0.5, an optimized threshold based on Youden's J statistic was applied to improve sensitivity and minimize missed infections. Each selected feature was then converted into a binary indicator using individual cut-points derived from ROC analysis, enabling construction of a straightforward cumulative score.

Feature-importance values were normalized to derive clinically interpretable weights. Highly influential predictors received slightly higher weights, while moderately informative features were scaled conservatively to balance performance with interpretability. The final scoring system was recalibrated using internal data and externally evaluated, demonstrating preserved sensitivity and specificity. This streamlined, transparent design ensures that the LDSS is suitable for routine use within laboratory workflows.

Validation of the LDSS
An independent validation dataset, obtained from an affiliated hospital within the same healthcare network, was used to assess the generalizability and robustness of the LDSS through temporal validation. This temporally separated retrospective dataset was entirely independent of all model development phases, including training, feature selection, and score construction.

Performance of the reduced 10-variable RF model and the three derived scoring systems was evaluated within this separate clinical environment. Standard classification metrics were computed and compared with those from the original external test set, providing insight into the system's real-world applicability.

The validation strategy adheres to recommendations from the International Federation of Clinical Chemistry and Laboratory Medicine for evaluating diagnostic tools using independent datasets. This approach strengthens the clinical credibility of the LDSS by demonstrating reproducibility across diverse healthcare settings.

Statistical Analysis
Descriptive statistics are presented as means ± standard deviations (SD) for continuous variables and as frequencies with percentages for categorical variables. Comparative analyses between the development and validation datasets were conducted using:

- Student's t-test for normally distributed continuous variables
- Welch's t-test for continuous variables with unequal variances or sample sizes
- Pearson's Chi-square test for categorical variables
- Z-tests for proportions and McNemar's test for paired categorical outcomes, particularly for comparing model performance metrics across datasets

These statistical comparisons were used to evaluate diagnostic consistency and identify significant differences in classification outcomes, providing insight into the reproducibility and robustness of the LDSS across diverse clinical settings.

All p-values were two-sided, with statistical significance defined as $p < 0.05$. Analyses were conducted using Python 3.13 and its associated statistical packages.

Results
Dataset Description and Data Preprocessing
The analytical cohort comprised 51,923 patient encounters, including 49,720 records from the main institutional database and 2,203 from an affiliated tertiary center. The validation cohort was enriched with inpatients from high-acuity units, such as Palliative Care and Gynecologic Oncology, and was specifically used to assess the external validity of the LDSS. The validation cohort demonstrated significantly higher age across all demographic strata (total: 43.92 vs. 38.28 years; males: 48.04 vs. 39.69; females: 41.23 vs. 37.41; all $p < 0.05$). Hematologic comparisons revealed statistically significant reductions in lymphocyte and eosinophil counts, accompanied by a modest but significant increase in HGB levels ($p < 0.05$).

Among urinalysis variables, the validation group exhibited higher bacterial counts, increased mucus presence, and elevated pH levels, whereas urine specific gravity and cylinder counts were lower ($p < 0.05$ for all). No significant differences were observed in WBC, neutrophil, monocyte, or basophil counts, nor in leukocyte counts, yeast presence, or gender distribution (all $p > 0.05$). Although the proportion of urine culture-positive cases was numerically similar (22.4% vs. 18.3%), this difference reached statistical significance ($p < 0.05$), potentially reflecting distinct microbiologic or clinical characteristics in the validation population.

Overall, these findings indicate that while the two datasets are broadly comparable, the validation cohort exhibits distinct demographic and laboratory profiles, likely due to its inpatient composition. These differences should be considered when interpreting LDSS performance in more complex clinical settings. Detailed summary statistics and p-values for each variable are provided in Table 1.

Hyperparameter Tuning
Each ML model was trained and optimized to achieve optimal performance on our dataset. Final hyperparameter configurations, tailored to the structure of each algorithm, are summarized in the Supplementary Material (Table S2).

Performance Metrics of ML Models
The performance of seven ML models was evaluated using both internal and external test datasets. Ensemble-based methods—RF, CatBoost, and XGBoost—consistently demonstrated high accuracy (≥0.929) and F1 scores (>0.83) across both datasets, highlighting their robustness for clinical prediction tasks.
On the external test set, RF outperformed all other models, achieving the highest ROC-AUC (0.956) and PR-AUC (0.907), indicating superior discrimination and precision-recall trade-off. CatBoost achieved the highest sensitivity (0.771) while maintaining balanced performance across other metrics.

KNN demonstrated exceptional specificity (0.988) and PPV (0.945) in the external set, making it particularly effective for ruling in cases. Conversely, LR, while computationally efficient, showed the lowest sensitivity and F1 scores, limiting its diagnostic utility.

Performance metrics from the external dataset closely mirrored those of the internal test set for all models, reinforcing their generalizability and stability. Comprehensive statistics for both datasets are provided in Table 2 and Figure 2. Among all evaluated algorithms, RF exhibited the most consistent and highest overall performance, with an internal ROC-AUC of 0.952 [95% CI: 0.948–0.956] and an external ROC-AUC of 0.956 [95% CI: 0.952–0.960], along with strong PR characteristics.

Given its superior accuracy, consistent generalizability, and interpretability, RF was selected as the core algorithm for integration into the LDSS. SHAP analysis was then performed on the final model to provide insight into the individual contribution of each feature to the predicted outcomes.

SHAP Analysis of the Optimal RF Model

Model interpretability was improved using SHAP, which quantifies the contribution of each feature to the predictions generated by the final RF model. As shown in Figure 3, the most influential features were

- Bacteria_Count (SHAP value: 0.061)
- Urine_Leu_Count (0.055)
- Nitrite (0.052)
- Age and Leukocyte Esterase (both 0.041)

These features correspond with well-established clinical markers of UTI, supporting the biological plausibility of the model.

Features with moderate importance included HGB, Gender, and Lymphocyte Count (LYM), with SHAP values ranging from 0.017 to 0.030. Features such as Bilirubin, Urobilinogen, and Ketone contributed minimally, each with SHAP values below 0.003.

Overall, the feature ranking confirms that the model primarily relies on clinically relevant variables, enhancing transparency and supporting its integration into laboratory decision-making.

Performance Metrics of the LDSS

A simplified RF model, built using the top 10 SHAP-derived features, maintained performance comparable to the full-feature model (ROC-AUC: 0.952 vs. 0.947; PR-AUC: 0.897 vs. 0.890), supporting its suitability for clinical implementation (Table 2). Based on these variables, three complementary scoring systems were developed to address distinct operational needs within laboratory workflows (Table 3):

- **Model-Prioritized Score:** Retains the behavior of the original machine-learning model by assigning weights directly from normalized SHAP values. This version is ideal for institutions seeking high overall discrimination while remaining faithful to the underlying algorithm.
- **Dual-Optimization Score:** Adjusts feature weights to balance sensitivity and specificity, as reflected in stable metrics across both test datasets (Table 4, Figure 4). This score is intended for laboratories aiming to minimize both missed infections and unnecessary cultures.
- **SAFE-Score:** Optimized for high sensitivity and NPV, this score is suitable for safety-critical settings where missing true infections is unacceptable—such as high-acuity units, elderly populations, or immunocompromised patients. Its higher sensitivity comes at the expense of specificity, highlighting the trade-off between diagnostic conservatism and resource utilization.

Across all scoring systems, sensitivity remained consistent in external and independent validation cohorts, while specificity varied according to prioritization strategy (Table 4). Together, these tools provide laboratories with flexible options that can be tailored to local clinical priorities, test-ordering practices, and antimicrobial stewardship goals (Figure 4).

Discussion

ML-based approaches offer substantial potential for the early diagnosis of UTIs. With the rising prevalence of antibiotic resistance, reducing unnecessary antibiotic use has become increasingly critical. Recent studies demonstrate that ML models improve diagnostic accuracy by integrating clinical symptoms, medical history, and urinary biomarkers, rather than relying solely on culture results [22].

Moreover, AI–driven decision-support systems (AI-DSS) can reduce diagnostic workload in hospitals, although their clinical validation remains limited [15]. Urinary biomarkers, such as nitrite and leukocyte esterase, exhibit high sensitivity for UTI diagnosis, yet their integration into ML models is essential to mitigate false-positive results [23]. AI-assisted methodologies are expected to be particularly beneficial for early detection of recurrent UTIs and multidrug-resistant pathogens, potentially improving patient outcomes and guiding more precise therapeutic interventions [23,24]. In this study, we evaluated the performance of multiple ML models in predicting urine culture outcomes and assessed their clinical applicability using explainable AI (XAI) techniques. Validation on a demographically and clinically distinct inpatient cohort further demonstrated the robustness and real-world adaptability of the LDSS. The incorporation of XAI enhanced interpretability, providing insight into the decision-making process and supporting potential integration in complex healthcare settings.

The LDSS was developed using all physician-ordered urine culture requests, including both culture-positive and culture-negative cases. Consequently, the dataset reflects the complete real-world distribution of suspected UTIs encountered in laboratory practice, enabling the model to learn discriminative patterns for both infection and non-infection samples.

Importantly, the LDSS functions solely as a laboratory-level decision-support tool rather than a diagnostic system. Its predictions are limited to variables available in the LIS and are intended to complement, not replace, physicians' diagnostic judgment.

Gender and Age-Related UTI Incidence

In our study, UTIs were significantly more common in female patients than in males. This finding aligns with existing literature and reinforces the well-established notion that women are more susceptible to UTIs due to urogenital anatomy, hormonal fluctuations, and lifestyle factors. Schmiemann et al. reported that UTI incidence in women is four to five times higher than in men [1]. Similarly, Hooton et al. identified a higher risk in women attributable to a shorter urethra and variability in periurethral microbial flora [25]. Additional risk factors include age, postmenopausal hormonal changes, and a history of recurrent infections.

Age also emerged as a critical determinant, with UTI incidence progressively increasing—particularly among women aged 65 years and older. While Foxman et al. reported peak incidence in women aged 15–29, with a secondary rise in postmenopausal groups [26], and Møller et al. linked estrogen depletion after age 50 to heightened susceptibility [11], our study identified older age (≥65 years) as an independent risk factor for positive urine culture in the LDSS model. This finding underscores the importance of incorporating age as a predictive variable and reflects the growing burden of UTIs in elderly populations.

Performance of ML Models

The predictive performance of the models developed in this study is consistent with, and in several cases surpasses, previously reported ML approaches for UTI prediction. Among the algorithms tested, ensemble-based models—particularly RF and CatBoost—demonstrated consistently high accuracy, balanced sensitivity and specificity, and favorable F1 scores. Compared to prior models reported by Vries et al. and Flores et al., our RF model showed superior performance across multiple evaluation metrics [2,27]. Likewise, our CatBoost implementation outperformed the model described by Mancini et al., which exhibited lower AUC and F1 values in a comparable clinical context [13].

Tree-based gradient boosting methods, such as XGBoost and LightGBM, also performed robustly and yielded results similar to high-performing models developed by Choi et al. and Lin et al., indicating strong generalizability across diverse patient populations [5,28]. In studies by Dhanda G et al. and Taylor RA et al., RF and XGBoost models similarly demonstrated superior discriminatory capacity, achieving AUC–ROC values of 0.85 and 0.90, respectively [29,30].

The KNN model achieved precision metrics comparable to prior studies; however, its limited interpretability may constrain clinical adoption [7]. Conversely, LR, while highly interpretable, exhibited lower sensitivity and F1 scores—consistent with Ramgopal et al., where the model tended to overpredict positive cases, reducing precision [10]. ANN (MLP) models, though commonly employed in UTI prediction studies, demonstrated moderate performance in our dataset, slightly below previously reported benchmarks [2].

Overall, these results reinforce the value of ensemble ML methods in the context of a LDSS for UTI prediction. They offer high predictive accuracy and consistent performance across internal and external validation cohorts, supporting their applicability in real-world clinical settings.

Several studies have investigated machine-learning–based urine culture prediction, varying in complexity and generalizability. Seheult et al. developed a decision-tree algorithm across multiple institutions to identify urinalysis predictors of culture positivity, reporting ROC-AUC values of approximately 0.78–0.79; however, their study lacked external validation and interpretability assessment. By comparison, our model achieved higher discrimination during development (ROC-AUC = 0.94–0.96) under cross-validation. Following conversion into a simplified score-based LDSS, real-world performance remained consistent (ROC-AUC ≈ 0.70–0.72; F1 ≈ 0.50–0.55). This decline reflects the expected trade-off between model complexity and clinical interpretability, as the LDSS was designed for practical integration into LIS rather than maximizing algorithmic precision [31].

Sergounioti et al. applied ensemble classifiers, including RF and XGBoost, to real-world laboratory data, achieving AUROC values of 0.79–0.82. However, their models combined clinical and laboratory parameters and lacked transparent feature-importance analysis. In contrast, our LDSS relied solely on structured laboratory data, achieved comparable discrimination (0.70–0.72), and preserved interpretability and reproducibility through rule-based score calibration via the Model-Prioritized and Dual-Optimization systems [32].

Sheele et al. investigated bacteriuria prediction in an emergency-department cohort using mixed clinical–laboratory features, yielding AUC–ROC values of 0.86–0.93 depending on the CFU/mL threshold. While their results were strong in a high-acuity population, our laboratory-only LDSS achieved comparable sensitivity (up to 95%) in routine diagnostic settings, highlighting its potential as a front-end decision-support tool for reflex culture testing [33].
Collectively, previous studies demonstrated the feasibility of ML-assisted urine culture prediction but often emphasized algorithmic performance over interpretability and clinical applicability. The present study addresses this gap by establishing a transparent, externally validated, and operational LDSS framework that maintains clinically acceptable performance while remaining fully interpretable and implementable within routine laboratory workflows.

Explainability and Feature Importance

SHAP-based feature-importance analysis in our study revealed a variable ranking that aligns with and extends existing literature. The most influential predictors were bacterial count, urine leukocyte count, nitrite, age, and leukocyte esterase. These findings are consistent with the meta-analysis by Devillé et al., which reported that combining nitrite and leukocyte esterase yielded a sensitivity of 88% and specificity of 98% for UTI diagnosis [8]. Similarly, Lachs et al. demonstrated that integrating these parameters with clinical symptoms significantly improves diagnostic accuracy [34].

Notably, our model also identified HGB levels, sex, and lymphocyte counts as important features with relatively high SHAP values, suggesting sensitivity to broader systemic or demographic factors that may influence infection risk. This aligns with Zhao Q et al., who reported age and sex among the top predictors in a SHAP-based post-urostomy UTI risk model [35], and Wang H et al., who found that systemic inflammatory markers and age were highly important in predicting post-surgical UTIs [36].

The predominance of microscopic urinalysis variables—particularly bacterial and leukocyte counts—over clinical or demographic features underscores the model's responsiveness to diagnostic biomarkers. This differentiates our approach from models such as Lee H et al., which focused on predicting antimicrobial resistance patterns but also leveraged SHAP analysis for interpretability [37].

Recent literature highlights the limitations of reflexive urine culture testing in the absence of clinical context. Munigala et al. and others have shown that reflex algorithms triggered by markers like leukocyte esterase or nitrite may reduce test volume but compromise diagnostic precision when symptom data are unavailable [38]. Fakih M et al. similarly argue that urinalysis alone is insufficient for accurate UTI diagnosis in asymptomatic patients, risking overdiagnosis and overtreatment [39].

Our study addresses the diagnostic gap through a reflective developed solely using structured laboratory data. Because symptom data are typically absent from LIS, the LDSS optimizes culture utilization within real-world laboratory constraints. Rather than functioning as an autonomous decision-maker or reflex trigger, the system serves as a reflective tool, providing SHAP-based analytical insights to support laboratory physicians' expert interpretation.
This reflective framework promotes standardized testing and interdisciplinary consultation. In equivocal cases, LDSS outputs can facilitate dialogue between laboratory and clinical teams, helping reconcile test reduction with diagnostic safety. Such an approach advances rational microbiological testing and provides a scalable model for clinician-laboratory collaboration [40].

The LDSS demonstrated robust predictive performance across internal and external datasets, supporting its seamless integration into routine laboratory workflows and reflective testing processes. The system is designed not to replace culture testing but to prioritize it based on evidence-driven probability, maintaining diagnostic stewardship.

To enhance accessibility for readers from diverse clinical and laboratory backgrounds, this study emphasizes the translational relevance of the LDSS over computational complexity. Its explainable design—supported by SHAP analysis and simplified scoring systems—enables non-technical users to interpret outputs transparently. While technical details were included to ensure methodological transparency and reproducibility, the interpretability of the system fosters trust, usability, and interdisciplinary communication between laboratory specialists and treating physicians. By promoting shared understanding of data-driven reasoning, the LDSS supports faster decision-making, improved test stewardship, and enhanced integration of laboratory insights into clinical workflows.

LDSS
Although symptom data were unavailable in the laboratory dataset, the LDSS was intentionally designed to function within the routine workflow of laboratory medicine, where test requests are frequently submitted without accompanying clinical narratives. By aligning the model with real-world laboratory constraints, the LDSS remains applicable and scalable across diverse clinical settings.

To improve interpretability and minimize unnecessary complexity, feature selection was applied to reduce the number of input variables. Prior studies have consistently demonstrated that parsimonious models are better suited for clinical implementation, as they are easier to interpret and maintain, while preserving acceptable predictive performance [41,42]. Accordingly, subsequent model development was restricted to ten key parameters that did not result in a statistically or clinically meaningful decline in performance. This strategy ensured an optimal balance between model simplicity and predictive accuracy.

Several published studies have similarly developed LDSS frameworks based on urine culture data, including those reported by de Vries et al., Dhanda et al., Del Ben et al., and Flores et al. [2,28,30,43]. Among these, Del Ben et al. employed a decision-tree-based approach, whereas the remaining studies selected RF as the primary algorithm [43]. The LDSS developed by de Vries and colleagues demonstrated performance metrics comparable to those observed in the present study, with AUC–ROC values ranging from 0.70 to 0.80. Although their model achieved a higher PPV, its NPV was lower than that of our model, highlighting differences in clinical trade-offs between false-positive and false-negative predictions.

Notably, Dhanda et al. and Flores et al. implemented scoring systems that stratified patients into high- and low-risk groups, an approach that is conceptually aligned with the strategy adopted in the present study. Across key performance metrics, the predictive accuracy of their models was broadly comparable to that of our system [2,30].

What distinguishes our LDSS is the integration of three distinct predictive models within a unified decision-making framework. To our knowledge, this is the first study to report the implementation of such a multi-model structure for UTI prediction. This design enables clinicians and laboratory physicians to select among alternative strategies according to specific clinical priorities, such as maximizing case detection or minimizing unnecessary diagnostic testing.

Although the SAFE-Score achieved excellent sensitivity, its specificity was limited (approximately 20%), a trade-off that may raise concerns regarding potential overtesting. Importantly, the LDSS was intentionally designed to accommodate this limitation by offering three complementary scoring strategies, each reflecting a distinct clinical philosophy. These include prioritization of patient safety (SAFE-Score), balanced diagnostic performance (Dual Optimization), and strict

adherence to model-derived predictions (Model-Prioritized). Rather than enforcing a one-size-fits-all solution, the LDSS functions as a flexible framework that facilitates consensus-based decision-making, allowing institutions to align model selection with local clinical expectations and operational priorities.

Crucially, the proposed system is not static. By continuously incorporating real-world data—particularly cases in which algorithmic recommendations are compared with expert laboratory physician judgments—the LDSS can be iteratively retrained and refined. As additional large-scale datasets are accumulated over time, improvements in specificity and overall diagnostic balance are anticipated, reflecting the inherent capacity of ML models to evolve with expanding data inputs. In this respect, the LDSS serves not only as an immediate decision-support tool but also as a scalable platform for continuous learning and performance optimization.

Within the Turkish healthcare context, reflective testing has not yet been systematically implemented. Nevertheless, the LDSS offers a structured and standardized framework that may facilitate its adoption, reduce inappropriate urine culture requests, and support antimicrobial stewardship initiatives. Moreover, the Ministry of Health of Türkiye has recently introduced a "Rational Laboratory Utilization" directive that explicitly promotes reflex and reflective testing practices. This regulatory emphasis is expected to accelerate the integration of reflective testing into routine laboratory workflows, highlighting the timeliness and practical relevance of the proposed system.

Finally, the LDSS was designed for seamless integration into routine clinical practice through Microsoft Excel, a widely available and familiar platform in most healthcare settings. All three predictive models are embedded within a single interface and generate concurrent outputs, enabling direct comparison and transparent interpretation at the point of use. Due to time constraints, the validation cohort was relatively small. Nevertheless, implementation of the LDSS within our hospital's central laboratory is planned, where it will be deployed to support real-time microbiological decision-making. This implementation will allow prospective validation of the system within routine laboratory workflows, evaluation of its diagnostic impact, and quantification of downstream outcomes, including reductions in unnecessary urine cultures, shorter turnaround times, and improved antibiotic stewardship. In addition, future multicenter studies across diverse healthcare systems are planned, incorporating structured clinical variables such as symptomatology, comorbidities, and medication history to further enhance the model's generalizability and clinical relevance.

Study Limitations

Although this study leveraged a large dataset and included external validation, several limitations should be acknowledged. First, all data were derived from a single healthcare network, which may limit generalizability to institutions with different patient populations, laboratory infrastructures, or clinical workflows. Second, the retrospective study design precluded assessment of the LDSS in real-time clinical decision-making; prospective Implementation studies are therefore required to determine its effects on clinical practice and patient outcomes.

Third, the model relied exclusively on structured laboratory data and did not incorporate patient symptoms, comorbidities, medication history, or clinical notes—factors known to influence UTI risk assessment and antibiotic prescribing. In routine clinical care, integration of such information is primarily the responsibility of the treating physician, who orders diagnostic tests based on patient history, clinical presentation, and prevailing guidelines. In contrast, laboratory physicians are tasked with processing submitted specimens according to standardized pre-analytical and analytical protocols. Although pre-preanalytical factors, such as appropriate test selection, are important, these data are rarely available to LIS in a structured, analyzable format. Consequently, most LIS environments contain only coded test orders and limited demographic information, without access to patient symptomatology or detailed clinical context.

Within these real-world constraints, the LDSS was designed not as a replacement for clinical judgment but as a complementary, interpretable decision-support tool that standardizes reflective testing and promotes communication between laboratory and clinical teams. Accordingly, the system functions as a laboratory-based reflex testing prioritization tool rather than as a diagnostic or therapeutic decision-making platform.

Fourth, despite robust performance in both internal and external test sets, the relatively small independent validation cohort—enriched for high-acuity inpatients—may introduce spectrum bias and lead to overestimation of sensitivity in complex clinical populations. Fifth, although the conventional definition of significant bacteriuria is $\geq 10^5$ CFU/mL, this study adopted a $\geq 10^4$ CFU/mL threshold based on emerging clinical evidence and institutional practice. Future investigations should evaluate the effects of alternative thresholds on model calibration and performance across different clinical settings.

Sixth, scoring weights and feature thresholds were calibrated using a fixed probability cutoff and Youden's index derived from the present dataset. Optimal thresholds may vary across institutions and will require local adjustment to maintain the desired balance between sensitivity and specificity. Finally, while SHAP values were employed to enhance model interpretability, clinician acceptance, usability, and integration into routine workflows were not formally assessed. Future implementation studies are therefore essential to evaluate user engagement, potential alert fatigue, and cost-effectiveness prior to widespread clinical deployment.

Conclusion

We developed and preliminarily validated an interpretable, multi-model LDSS designed to improve the efficiency of urine culture utilization. By integrating ensemble machine-learning approaches with SHAP-based interpretability, the system demonstrated strong discriminatory performance while offering flexible scoring strategies that prioritize sensitivity, specificity, or an optimized balance between the two. The LDSS has the potential to reduce unnecessary urine cultures, support antimicrobial stewardship efforts, and promote standardized, evidence-based laboratory decision-making. Future work will focus on prospective, real-world implementation across diverse clinical settings. Planned enhancements include integration with electronic health record–derived clinical data, local calibration of decision thresholds, and

systematic evaluation of clinical impact, user adoption, and cost-effectiveness. These steps are critical for translating this early-stage model into a scalable and clinically actionable decision-support tool.

Figure 1. STARD flow diagram of study participants and urine culture testing.
Figure 2. Receiver operating characteristic (ROC) and precision-recall (PR) curves illustrating the predictive performance of ML models.
Figure 3. SHAP summary plot showing variable importance in the RF model.
Figure 4. LDSS workflow illustrating selection criteria based on diagnostic accuracy and operational priorities.

Table 1. Baseline characteristics of the study population, including demographic, clinical, and laboratory variables.

Characteristics [a]	Unit	Main Dataset (n = 49,720) Mean ± SD	Training Set (n = 29,832) Mean ± SD	Internal Test Set (n = 9,944) Mean ± SD	External Test Set (n = 9,944) Mean ± SD	Validation Set (n = 2,203) Mean ± SD	p-value [b] (Main Dataset vs Validation Set)
Age		38.28 ± 26.85	38.07 ± 26.81	38.89 ± 26.99	38.29 ± 26.83	43.92 ± 28.53	<0.05
Male	Years	39.69 ± 28.20	39.33 ± 28.12	40.09 ± 28.39	40.37 ± 28.26	48.04 ± 28.38	<0.05
Female	Years	37.41 ± 25.96	37.29 ± 25.95	38.17 ± 26.07	37.03 ± 25.85	41.23 ± 28.33	<0.05
Gender							0.152
Male	n (%)	18,871 (38.0%)	11,358 (38.1%)	3,766 (37.9%)	3,747 (37.7%)	870 (39.5%)	
Female	n (%)	30,849 (62.0%)	18,474 (61.9%)	6,178 (62.1%)	6,197 (62.3%)	1333 (60.5%)	
WBC	$\times 10^9$ cells/L	8.47 ± 4.63	8.5 ± 4.91	8.4 ± 3.86	8.45 ± 4.47	8.45 ± 3.48	0795
Neutrophil	$\times 10^9$ cells/L	5.1 ± 3.4	5.11 ± 3.34	5.05 ± 3.13	5.11 ± 3.81	5.18 ± 3.14	0244
Lymphocyte	$\times 10^9$ cells/L	2.45 ± 2.82	2.47 ± 3.27	2.42 ± 1.97	2.43 ± 1.99	2.36 ± 1.26	<0.05
Monocyte	$\times 10^9$ cells/L	0.68 ± 0.85	0.68 ± 1.01	0.68 ± 0.68	0.67 ± 0.37	0.67 ± 0.29	0.168
Eosinophil	$\times 10^9$ cells/L	0.2 ± 0.25	0.2 ± 0.25	0.2 ± 0.25	0.2 ± 0.24	0.19 ± 0.19	<0.05
Basophil	$\times 10^9$ cells/L	0.04 ± 0.06	0.04 ± 0.06	0.03 ± 0.05	0.04 ± 0.07	0.04 ± 0.03	1.000
HGB	g/dL	12.26 ± 1.91	12.26 ± 1.9	12.27 ± 1.92	12.27 ± 1.91	12.56 ± 1.98	<0.05
Bacteria Count (urine)	/HPF	33.57 ± 124.45	33.55 ± 127.66	33.93 ± 120.89	33.24 ± 118.07	41.7 ± 157.49	<0.05
Leucocyte Count (urine)	/HPF	53.46 ± 287.59	53.81 ± 288.38	52.32 ± 279.64	53.53 ± 293.02	64.28 ± 324.2	0.124
Yeast	/HPF	3.85 ± 133.83	5.04 ± 170.15	1.95 ± 36.1	2.2 ± 37.23	3.13 ± 55.43	0.587
Mucus	/HPF	11.32 ± 30.73	11.34 ± 30.69	10.97 ± 28.36	11.62 ± 33.03	22.14 ± 56.43	<0.05
Cylinder	/HPF	0.04 ± 0.22	0.04 ± 0.22	0.04 ± 0.23	0.05 ± 0.23	0 ± 0	<0.05
Density	-	1016.98 ± 8.17	1017.02 ± 8.14	1016.95 ± 8.23	1016.9 ± 8.22	1015.86 ± 7.19	<0.05
pH	-	5.9 ± 0.81	5.91 ± 0.82	5.89 ± 0.81	5.9 ± 0.81	6.05 ± 0.52	<0.05
Urine Culture							<0.05
Positive	n	11,156 (22.4%)	6,694 (22.4%)	2,231 (22.4%)	2,231 (22.4%)	403 (18.3%)	1.000
Negative	n	38,564 (77.6%)	23,138 (77.6%)	7,713 (77.6%)	7,713 (77.6%)	1,800 (81.7%)	1.000

a Categorical variables were not included in this table.
b Continuous variables were compared using Welch's t-test, and categorical variables were analyzed with Pearson's chi-square test. A p-value < 0.05 was considered statistically significant.

Table 2. Classification performance metrics of the ML models, including accuracy, sensitivity, specificity, and AUC.

Model	Sensitivity	Specificity	PPV	NPV	Accuracy	F1 Score	ROC-AUC	PR-AUC
Internal Test Set								

RF	0.758 (0.741–0.776)	0.985 (0.982–0.987)	0.934 (0.923–0.946)	0.934 (0.929–0.939)	0.934 (0.929–0.938)	0.838 (0.826–0.850)	0.952 (0.948–0.956)	0.897 (0.891–0.903)
XGBoost	0.768 (0.751–0.784)	0.976 (0.973–0.979)	0.902 (0.889–0.916)	0.936 (0.930–0.941)	0.929 (0.925–0.934)	0.830 (0.816–0.842)	0.930 (0.925–0.935)	0.861 (0.854–0.868)
LightGBM	0.681 (0.664–0.699)	0.972 (0.968–0.976)	0.876 (0.862–0.894)	0.913 (0.907–0.919)	0.907 (0.900–0.913)	0.766 (0.751–0.780)	0.916 (0.911–0.921)	0.825 (0.818–0.832)
CatBoost	0.764 (0.747–0.784)	0.980 (0.977–0.983)	0.918 (0.907–0.931)	0.935 (0.930–0.940)	0.932 (0.927–0.937)	0.834 (0.822–0.847)	0.930 (0.925–0.935)	0.861 (0.854–0.868)
Logistic Regression	0.350 (0.330–0.370)	0.969 (0.965–0.973)	0.765 (0.738–0.791)	0.838 (0.830–0.846)	0.830 (0.823–0.837)	0.480 (0.459–0.501)	0.790 (0.782–0.798)	0.593 (0.583–0.603)
ANN (MLP)	0.561 (0.541–0.582)	0.943 (0.937–0.947)	0.738 (0.717–0.758)	0.881 (0.875–0.888)	0.857 (0.850–0.864)	0.637 (0.621–0.655)	0.844 (0.837–0.851)	0.698 (0.689–0.707)
KNN	0.723 (0.705–0.743)	0.984 (0.981–0.987)	0.929 (0.917–0.940)	0.925 (0.919–0.930)	0.925 (0.920–0.931)	0.813 (0.801–0.827)	0.947 (0.943–0.951)	0.903 (0.897–0.909)
RF (with top 10 variables) *	0.769 (0.761–0.777)	0.981 (0.979–0.984)	0.924 (0.919–0.930)	0.936 (0.931–0.941)	0.934 (0.929–0.939)	0.8397 (0.832–0.847)	0.947 (0.944–0.952)	0.890 (0.884–0.896)
External Test Set								
RF	0.76 (0.744–0.778)	0.987 (0.984–0.989)	0.943 (0.932–0.953)	0.935 (0.929–0.94)	0.936 (0.931–0.941)	0.842 (0.829–0.854)	0.956 (0.952–0.96)	0.907 (0.901–0.913)
XGBoost	0.767 (0.748–0.784)	0.980 (0.977–0.983)	0.917 (0.906–0.930)	0.936 (0.930–0.942)	0.932 (0.928–0.938)	0.836 (0.824–0.848)	0.932 (0.927–0.937)	0.877 (0.871–0.883)
LightGBM	0.686 (0.666–0.704)	0.976 (0.972–0.979)	0.892 (0.877–0.907)	0.915 (0.909–0.921)	0.911 (0.905–0.916)	0.776 (0.762–0.789)	0.919 (0.914–0.924)	0.840 (0.833–0.847)
CatBoost	0.771 (0.754–0.790)	0.982 (0.979–0.985)	0.924 (0.911–0.936)	0.936 (0.931–0.942)	0.934 (0.929–0.939)	0.840 (0.827–0.852)	0.929 (0.924–0.934)	0.875 (0.868–0.882)
Logistic Regression	0.339 (0.321–0.358)	0.968 (0.964–0.972)	0.755 (0.725–0.781)	0.835 (0.828–0.842)	0.827 (0.819–0.834)	0.467 (0.445–0.487)	0.793 (0.785–0.801)	0.597 (0.587–0.607)
ANN (MLP)	0.565 (0.544–0.585)	0.937 (0.932–0.943)	0.722 (0.700–0.744)	0.881 (0.874–0.888)	0.854 (0.847–0.861)	0.634 (0.618–0.651)	0.846 (0.839–0.853)	0.707 (0.698–0.716)
KNN	0.719 (0.700–0.738)	0.988 (0.985–0.990)	0.945 (0.933–0.955)	0.924 (0.918–0.929)	0.927 (0.923–0.933)	0.817 (0.803–0.830)	0.947 (0.943–0.951)	0.905 (0.899–0.911)

*Reduced model including only the top 10 predictors selected by SHAP analysis: bacterial count in urine, urinary leukocyte count, urinary nitrite test, patient age, leukocyte esterase activity in urine, HGB concentration, gender, lymphocyte count, urine density, and urinary erythrocyte count.

Table 3A. Confusion matrix–derived performance metrics of the ML models, including sensitivity, specificity, PPV, and NPV.

Feature	Threshold Binarization	Normalized SHAP Value	Model-Prioritized Score System [1]	Dual-Optimization Score[2]	SAFE-Score System (Sensitive Assessment for Exclusion)[3]	Scientific Justification
Bacteria Count	>20	0.175	0.20	0.32	0.89	Major diagnostic marker for infection; emphasized clinically.
Urine Leukocyte Count	>25	0.157	0.18	0.22	0.05	Strongly correlates with infection; slightly boosted for sensitivity.
Nitrite	= 1	0.147	0.17	0.15	0.77	Positive nitrite is a direct indicator of gram-negative

						bacterial activity.
Age	>65	0.118	0.15	0.23	0.42	Increased risk in elderly population (>65 years).
Leucocyte Esterase	>0	0.116	0.14	0.13	0.82	Biochemical indicator of leukocytes; moderate importance.
HGB	<12	0.085	0.10	0.12	0.71	Low HGB levels linked to increased infection susceptibility.
Gender	= 1 (Female)	0.062	0.08	0.04	0.06	Higher infection prevalence anatomically in females.
LYM	<1.5	0.051	0.06	0.1	0.65	Low lymphocyte count indicates immunosuppression risk.
Density	>1020	0.048	0.05	0.09	0.03	Higher urine density occasionally correlates with infection.
Urine Erythrocyte	>0	0.047	0.05	0.1	0.77	Presence may suggest urinary tract pathology but less specific.

1. The first system was developed using model-derived, data-driven thresholds and weighting.
2. The second system was designed to optimize both sensitivity and specificity, achieving balanced classification performance.
3. The third system prioritized minimizing false negatives, emphasizing maximum sensitivity and PNV.

Table 4. Performance metrics of the LDSS evaluated using both external test and validation datasets.
A. Results from the external test set.

Method	Sensitivity	Specificity	PPV	NPV	PLR	NLR	Accuracy	F1 Score	ROC-AUC	PR-AUC
Model-Prioritized Score System[1]	55.94 (53.87–57.99)	85.83 (85.03–86.59)	53.31 (51.29–55.32)	87.07 (86.30–87.81)	3.95 (3.69–4.22)	0.51 (0.49–0.54)	79.1 (78.31–79.91)	54.59 (52.57–56.60)	70.88 (67.58–74.28)	54.62 (50.71–57.71)
Dual-Optimization Score System[2]	64.77 (62.76–66.72)	76.62% (75.67–77.55)	44.49 (42.79–46.20)	88.26 (87.47–89.01)	2.77 (2.47–3.07)	0.46 (0.44–0.49)	73.96 (73.09–74.82)	52.75 (51.03–54.46)	70.70 (68.70–72.70)	54.63 (52.72–56.72)
SAFE-Score System[3]	95.34 (94.38–96.14)	20.29% (19.41–21.20)	25.70 (24.77–26.66)	93.77 (92.51–94.83)	1.20 (1.03–1.136)	0.23 (0.21–0.25)	37.13 (36.18–38.08)	40.49 (39.44–41.55)	57.81 (57.81–57.81)	60.52 (55.55–65.58)

1. TP = 1,248; TN = 6,620; FP = 1,093; FN = 983
2. TP = 1,445; TN = 5,910; FP = 1,803; FN = 786
3. TP = 2,127; TN = 1,565; FP = 6,148; FN = 104

B. Results from the validation test set.

Method	Sensitivity	Specificity	PPV	NPV	PLR	NLR	Accuracy	F1 Score	ROC-AUC	PR-AUC
Model-Prioritized Score System[1]	57.95 (53.00–62.78)	84.78 (83.04–86.41)	46.47 (43.09–49.89)	89.87 (88.74–90.85)	3.81 (3.32–4.37)	0.50 (0.44–0.56)	79.80 (78.06–81.46)	51.51 (47.30–55.37)	71.31 (68.60–73.94)	34.80 (31.05–38.61)
Dual-Optimization Score System[2]	66.50 (61.70–71.07)	76.48 (74.44–78.42)	39.19 (36.65–41.80)	90.92 (89.71–92.00)	2.83 (2.54–3.15)	0.44 (0.38–0.50)	74.63 (72.75–76.43)	49.36 (45.86–53.01)	71.40 (68.82–73.94)	32.35 (29.00–35.66)
SAFE-Score System[3]	94.87 (92.26–96.79)	24.30 (22.33–26.36)	22.22 (21.63–22.83)	95.40 (93.14–96.95)	1.25 (1.21–.1.30)	0.21 (0,14–0,32)	37.40 (35,38–39.46)	36.08 (33.46–38.63)	59.58 (58.14–61.03)	22.03 (20.05–24.01)

1. TP = 237; TN = 1,521; FP = 273; FN = 172
2. TP = 272; TN = 1,372; FP = 422; FN = 137
3. TP = 388; TN = 436; FP = 1,358; FN = 21