Article

Not peer-reviewed version

From Preliminary Urinalysis to Decision Support: Machine Learning for UTI Prediction in Real-World Laboratory Data

Athanasia Sergounioti * , Dimitrios Rigas , Vassilios Zoitopoulos , Dimitrios Kalles

Posted Date: 1 April 2025

doi: 10.20944/preprints202504.0037.v1

Keywords: urinary tract infection; machine learning; urinalysis; antimicrobial stewardship; threshold optimization; XGBoost; diagnostic decision support

 doi:10.20944/preprints202504.0037.v1

Article

From Preliminary Urinalysis to Decision Support: Machine Learning for UTI Prediction in Real-World Laboratory Data

Athanasia Sergounioti [1,*], Dimitrios Rigas [2], Vassilios Zoitopoulos [1] and Dimitrios Kalles [3]

[1] Department of Laboratory Medicine, General Hospital of Amfissa, Greece; nasiaser@yahoo.gr, miclab1@gnamfissas.gr

[2] Independent Researcher; rigas_dimitris@yahoo.gr

[3] School of Science and Technology, Hellenic Open University, 26335 Patras, Greece; kalles@eap.gr

* Correspondence: nasiaser@yahoo.gr

Abstract: Background/Objectives: Urinary tract infections (UTIs) are frequently diagnosed empirically, often leading to overtreatment and rising antimicrobial resistance. This study aimed to develop and evaluate machine learning (ML) models that predict urine culture outcomes using routine urinalysis and demographic data, supporting more targeted empirical antibiotic use. **Methods:** A real-world dataset comprising 8,065 urinalysis records from a hospital laboratory was used to train five ensemble ML models: Random Forest, XGBoost (eXtreme Gradient Boosting), Extra Trees, Voting Classifier, and Stacking Classifier. Models were developed using 10-fold stratified cross-validation and assessed via clinically relevant metrics including specificity, sensitivity, likelihood ratios, and diagnostic odds ratio (DOR). To enhance screening utility, threshold optimization was applied to the best-performing model (XGBoost) using the Youden index. **Results:** XGBoost and Random Forest demonstrated the most balanced diagnostic profiles, with DORs exceeding 21. The Voting and Stacking Classifiers achieved highest specificity (>95%) and positive likelihood ratios (>10), but exhibited lower sensitivity. Feature importance analysis identified positive nitrites, white blood cell count, and specific gravity as key predictors. Threshold tuning of XGBoost improved sensitivity from 70.2% to 87.9% and reduced false negatives by 82%, with an associated NPV of 96.4%. The adjusted model reduced overtreatment by 56% compared to empirical prescribing. **Conclusions:** ML models based on structured urinalysis and demographic data can support clinical decision-making for UTIs. While high-specificity models may reduce unnecessary antibiotic use, sensitivity trade-offs must be considered. Threshold-optimized XGBoost offers a clinically adaptable tool for empirical treatment decisions, particularly in settings lacking rapid diagnostics.

Keywords: urinary tract infection; machine learning; urinalysis; antimicrobial stewardship; threshold optimization; XGBoost; diagnostic decision support

1. Introduction

Urinary tract infections (UTIs) are among the most prevalent infections worldwide, posing a significant clinical and economic burden on healthcare systems [1]. They represent the most common outpatient infections, with at least half of all adult women experiencing a UTI during their lifetime. In healthcare settings, UTIs account for up to 9.4% of all hospital-acquired infections [2]. Despite advancements in diagnostics and treatment, UTIs continue to be associated with high morbidity and, in some cases, mortality [3].

Diagnosing UTIs remains challenging due to nonspecific symptoms and the delay in obtaining urine culture results. In emergency and primary care settings, clinicians often initiate empirical antibiotic treatment based on clinical suspicion to avoid delays in managing potentially severe infections [4]. However, this practice contributes to over-diagnosis and antibiotic overuse,

 doi:10.20944/preprints202504.0037.v1

contributing to the development of antimicrobial resistance. Previous studies suggest that 40–50% of patients empirically treated for UTI may not have a true infection [5] and broader estimates indicate that approximately 28% of all antibiotic prescriptions may be unnecessary [6].

Machine learning (ML) offers a data-driven approach to enhance diagnostic accuracy and support antimicrobial stewardship. Recent work has explored ML-based prediction of urine culture outcomes using standard urinalysis data. For example, Parente et al. developed a random forest model to identify low-risk patients in primary care, potentially reducing unnecessary antibiotic use [7]. Dedeene et al. proposed an artificial intelligence support tool for rapid prediction of culture results in emergency settings [8] while Seheult et al. implemented a decision tree model for optimizing urinalysis-based UTI prediction [9].

Additional studies have explored related approaches, including the prediction of recurrent UTIs in pediatric populations using ML algorithms [10] and the development of interpretable models for critical UTI outcomes [11]. Zhang et al. proposed an AI-based prediction framework for both urinary tract infections and associated bloodstream infections, supporting early identification of high-risk patients [12]. Recent advances have also emphasized the importance of explainability in model development, facilitating clinical trust and adoption [13]. In a recent conceptual model, point-of-care diagnostic strategies for UTIs were shown to support antimicrobial stewardship by reducing empirical treatment in low-risk patients [14]. Recent findings suggest that the predictive performance of urinalysis may vary depending on the specific causative microorganism, indicating a need for microorganism-aware modeling strategies [15].

These advances highlight the potential of machine learning to improve diagnostic decision-making in UTI care. The present study aims to evaluate and compare the performance of five supervised ensemble ML models in predicting UTI based on urinalysis and demographic features. Emphasis is placed on diagnostic metrics with clinical relevance—including specificity, positive predictive value (PPV), and positive likelihood ratio (PLR)—with the goal of supporting more rational empirical antibiotic use.

The key contributions of this study include: (i) development of an interpretable ML pipeline using real-world laboratory data; (ii) comparative performance analysis of five ensemble classifiers; and (iii) proposal of high-specificity models as potential tools to improve the management of empirical antibiotic prescribing in suspected UTIs—thereby supporting the implementation of a personalized approach in managing UTIs.

2. Materials and Methods

The Materials and Methods should be described with sufficient details to allow others to replicate and build on the published results. Please note that the publication of your manuscript implicates that you must make all materials, data, computer code, and protocols associated with the publication available to readers. Please disclose at the submission stage any restrictions on the availability of materials or information. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited.

2.1. Dataset Description

The dataset used in this study was collected from the Laboratory Medicine Department of the General Hospital of Amfissa, Greece. It comprised urinalysis parameters, urine culture results, and demographic information from patients who underwent routine urinalysis and urine culture. Urinalysis data were obtained using the iChemVELOCITY Automated Urinalysis Analyzer (Leriva Diagnostics), an automated laboratory instrument employed in routine clinical practice. The analyzer evaluates urine dipstick tests using reflectance photometry and standardized colorimetric methods, providing operator-independent measurements of physical and chemical urine characteristics. The reported parameters included specific gravity (SG), pH, color, clarity, ketones, glucose, protein, nitrites, bilirubin, urobilinogen, white blood cells (WBC), and blood. All results were retrospectively extracted from the laboratory information system and fully de-identified prior to analysis. Urine

 doi:10.20944/preprints202504.0037.v1

culture results were classified as positive when bacterial growth exceeded 10^5 cfu/mL, and as non-positive otherwise, in accordance with established microbial threshold guidelines [16-18].

The dataset included 8,065 urine samples, of which 2,257 (28.0%) were classified as positive and 5,808 (72.0%) as non-positive. Of the total samples, 3,111 (38.6%) were from male and 4,954 (61.4%) from female patients. The mean patient age was 57.8 years (SD = 25.0). Among males, 22.6% of cultures were positive, while the positivity rate among females was 29.1%.

2.2. Data Preprocessing

- Missing Data Handling: Entries with incomplete laboratory values were excluded from the analysis. No imputation techniques were applied, in order to preserve the integrity and original distribution of the data.
- Feature Engineering and Encoding: Categorical variables (gender, color, clarity) were encoded using one-hot or ordinal encoding, depending on their characteristics. Numerical variables were standardized where appropriate [19]. All available variables were retained across models to ensure comparability, and no dimensionality reduction techniques were applied.
- Class Imbalance Handling: Although the proportion of positive cultures (26.2%) did not constitute severe imbalance, it was sufficient to potentially bias model learning. Initial attempts using synthetic oversampling techniques such as SMOTE and ADASYN degraded precision and increased false positives due to noisy synthetic samples. Therefore, all models were ultimately trained using a Balanced Bagging Classifier framework. This ensemble-based resampling method improves minority class representation while preserving the original data structure and avoiding synthetic artifacts. Its reliability in structured clinical datasets has been previously demonstrated [20-22].

2.3. Model Development

This study focused on five ensemble learning algorithms, selected for their robustness, generalization capacity, and proven performance in clinical tabular data: Random Forest, XGBoost, Extra Trees, Voting Classifier, and Stacking Classifier. Each model was developed using 10-fold stratified cross-validation to ensure balanced representation of positive and negative cases across training and validation subsets. All preprocessing steps—including encoding and scaling—were embedded within scikit-learn pipelines to prevent data leakage [23-26].

The Random Forest and Extra Trees classifiers are both ensembles of decision trees but differ in how splits are determined; Extra Trees uses full randomization to reduce variance and is often less prone to overfitting in imbalanced settings. XGBoost, a gradient-boosted decision tree algorithm optimized for predictive accuracy, was fine-tuned using randomized grid search over key hyperparameters such as learning rate, maximum depth, and number of estimators [27-28].

The Voting Classifier aggregates predictions from Random Forest, XGBoost, and Extra Trees using soft voting, averaging the predicted class probabilities to determine the final outcome. The Stacking Classifier employs the same base learners, with Logistic Regression as the meta-learner to generate a final calibrated prediction. This setup offers the advantage of combining model diversity with interpretable outputs [29-32].

All models were implemented using the scikit-learn and XGBoost libraries in Python. Preprocessing and model training were structured using scikit-learn's Pipeline and ColumnTransformer objects. The ColumnTransformer allowed for the application of different preprocessing steps to numeric and categorical features, whereas the Pipeline combined preprocessing and modeling into a single, cross-validation-safe workflow, preventing information leakage and ensuring reproducibility.

 doi:10.20944/preprints202504.0037.v1

2.4. Evaluation Metrics

Performance was evaluated using a set of standard diagnostic metrics, selected to provide a structured and clinically relevant assessment of model performance:

- Accuracy: Proportion of correct predictions (true positives and true negatives) among all evaluated cases.
- Balanced Accuracy: Mean of sensitivity and specificity. Suitable for imbalanced datasets as it considers both classes equally.
- Sensitivity (Recall): Proportion of actual positive cases correctly identified by the model.
- Specificity: Proportion of actual negative cases correctly identified. High specificity reduces the likelihood of false positives.
- Precision (Positive Predictive Value - PPV): Proportion of predicted positive cases that are true positives.
- Negative Predictive Value (NPV): Proportion of predicted negative cases that are true negatives.
- F1-score: Harmonic mean of precision and recall. Useful when both false positives and false negatives carry clinical consequences.
- ROC AUC: Area under the Receiver Operating Characteristic curve. Reflects overall discrimination capacity of the model across thresholds.
- Matthews Correlation Coefficient (): A balanced measure that incorporates all components of the confusion matrix. Appropriate for imbalanced datasets.
- Positive Likelihood Ratio (PLR): Indicates how much more likely a positive test result is in someone with the condition than in someone without it. Values >10 are considered strong evidence to support a diagnosis.
- Negative Likelihood Ratio (NLR): Ratio of the false negative rate to the true negative rate. Lower values suggest the test is effective at excluding disease; values <0.1 are generally considered acceptable.
- Diagnostic Odds Ratio (DOR): The ratio of the odds of a positive test result in patients with the disease to the odds of the same result in those without it. It integrates sensitivity and specificity into a single indicator of discriminative power [33-35].

2.5. Model Interpretation and Performance Visualization

To assess model interpretability, we applied Permutation Feature Importance (PFI), a model-agnostic approach that quantifies the contribution of each predictor by measuring the change in model performance when the values of that feature are randomly shuffled. This method was implemented using the permutation importance function from scikit-learn, applied to cross-validated predictions to ensure robustness [36-37].

In addition to importance analysis, ROC curves, precision-recall curves, and confusion matrices were generated to support graphical evaluation of diagnostic performance. These plots were created using predicted probabilities and model outputs aggregated across all cross-validation folds. Interpretation and visualization focused primarily on the top-performing models, as identified by their high specificity, PPV, and positive likelihood ratios.

2.6. Threshold Optimization

Among the evaluated classifiers, the XGBoost model demonstrated the most clinically balanced performance in terms of diagnostic accuracy, interpretability, and generalization. Owing to its favorable diagnostic odds ratio and high ROC AUC, it was selected for further optimization. To improve its clinical sensitivity, we applied probability threshold tuning by analyzing performance metrics across a range of classification cutoffs. The optimal threshold was determined using the Youden index [38], with the objective of enhancing the model's ability to detect true positive cases

 doi:10.20944/preprints202504.0037.v1

without excessively compromising specificity. This post-training adjustment aimed to minimize false negatives—an important consideration in the early management of urinary tract infections.

3. Results

The comparative performance of all models is summarized in Table 1. While the Voting and Stacking Classifiers achieved the highest specificity (95.6% and 95.4%, respectively) and positive likelihood ratios (PLR >10), their sensitivity was notably low, indicating limited ability to detect all true positive cases. These characteristics suggest potential utility in scenarios where false positives must be minimized, but raise concerns regarding their suitability for general screening purposes.

In contrast, the XGBoost classifier exhibited the most balanced overall profile. Combining high specificity with strong overall discriminative performance, it emerged as a versatile model for clinical settings— supporting the exclusion of culture-negative cases without disproportionately compromising detection of true infections. The Random Forest classifier showed similar behavior, with a closely aligned diagnostic odds ratio and performance trade-offs. These two models offered a more even distribution of diagnostic strengths, suggesting their greater suitability for scenarios where both false positives and false negatives carry clinical implications.

The Extra Trees classifier also performed well in terms of specificity and positive likelihood ratio, although its overall detection capacity appeared more limited. This places it between the high-specificity models and those offering broader balance across diagnostic metrics.

Across all classifiers, the negative predictive value (NPV) remained high (>88%), suggesting reliability in identifying patients without UTI. However, none of the models achieved an NLR below 0.1, a threshold typically required for effective rule-out tools. As such, while models like XGBoost may assist in reducing unnecessary empirical treatment, they should be interpreted cautiously and within a broader clinical decision-making framework.

Table 1. presents the detailed results, including accuracy, sensitivity, specificity, PPV, NPV, F1-score, balanced accuracy, PLR, NLR, and DOR for each classifier.

Model	Accuracy	Balanced Accuracy	Precision (PPV)	Recall (Sensitivity)	Specificity	NPV	F1-score	ROC AUC	MCC	PLR	NLR	DOR
Random Forest	0.856	0.808	0.596	0.731	0.885	0.934	0.657	0.888	0.572	6.42	0.30	21.65
XGBoost	0.859	0.799	0.610	0.702	0.896	0.928	0.670	0.892	0.575	6.79	0.33	21.20
Extra Trees	0.857	0.756	0.626	0.595	0.918	0.918	0.655	0.888	0.571	7.34	0.44	17.23
Stacking Classifier	0.864	0.714	0.706	0.473	0.954	0.887	0.566	0.883	0.503	10.69	0.552	19.77
Voting Classifier	0.868	0.721	0.717	0.486	0.956	0.889	0.579	0.891	0.518	11.29	0.538	21.63

Table 1: Diagnostic performance metrics of five ensemble classifiers for UTI prediction using urinalysis and demographic features.

For all models, feature importance was assessed using permutation importance (PFI) based on accuracy loss. Feature ranking was visualized using bar plots, and consistent key predictors across models included WBC count, protein, and urine pH.

In addition to importance analysis, ROC curves (Figure 1) and precision–recall curves (Figure 2) were generated to support graphical evaluation of diagnostic performance. These plots were created using predicted probabilities and model outputs aggregated across all cross-validation folds. To further enhance model interpretability, permutation feature importance (PFI) was applied to the XGBoost and Random Forest classifiers. As illustrated in Figure 3, both models consistently identified

 doi:10.20944/preprints202504.0037.v1

positive nitrites test (NITRITES_POSITIVE), WBC count, and specific gravity (SG) as the most informative predictors for UTI diagnosis.

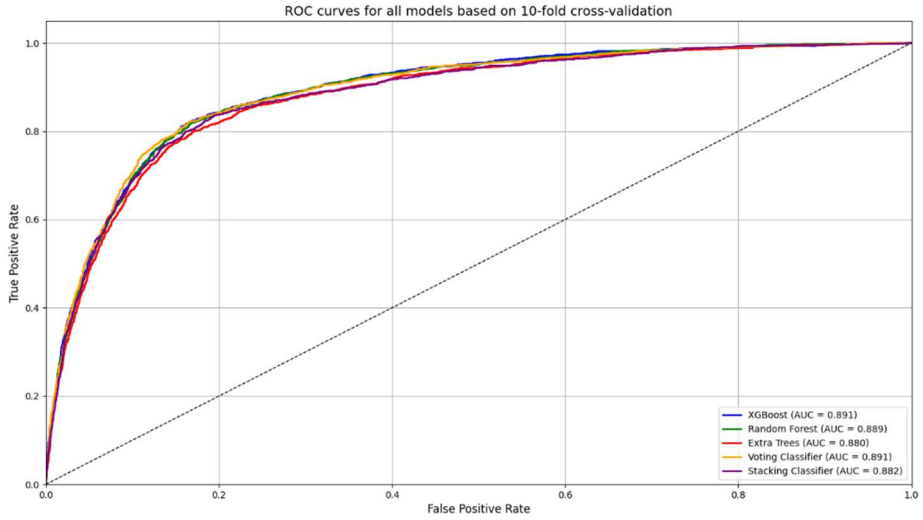

Figure 1. ROC curves for all models based on 10-fold cross-validation. All classifiers demonstrated comparable AUC values, with XGBoost achieving the highest value (AUC = 0.892).

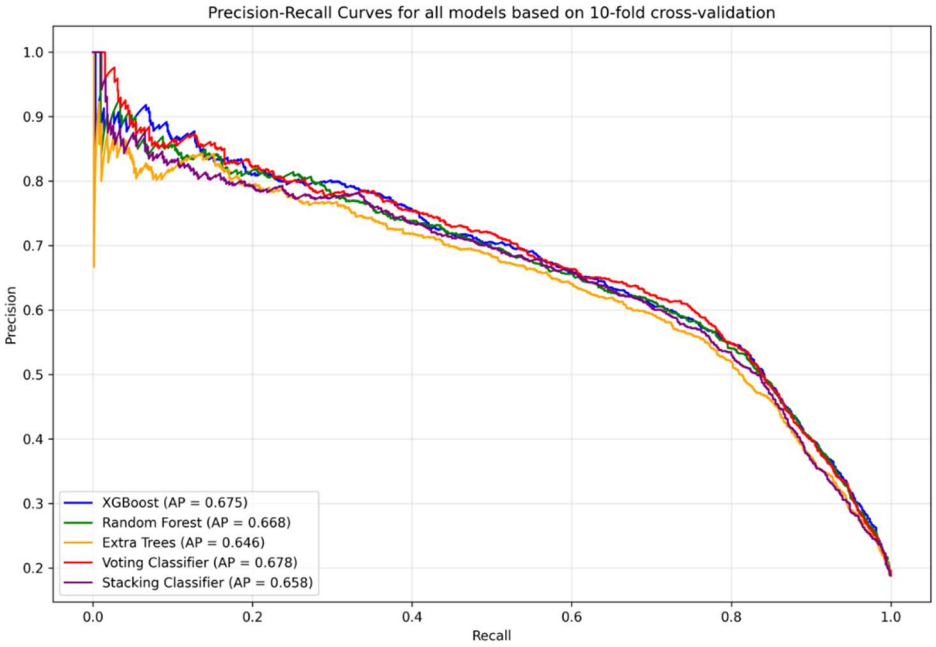

Figure 2. Precision–recall curves for all models based on 10-fold stratified cross-validation. Despite their higher average precision, the Voting and Stacking classifiers exhibited markedly reduced recall, limiting their use in general screening.

 doi:10.20944/preprints202504.0037.v1

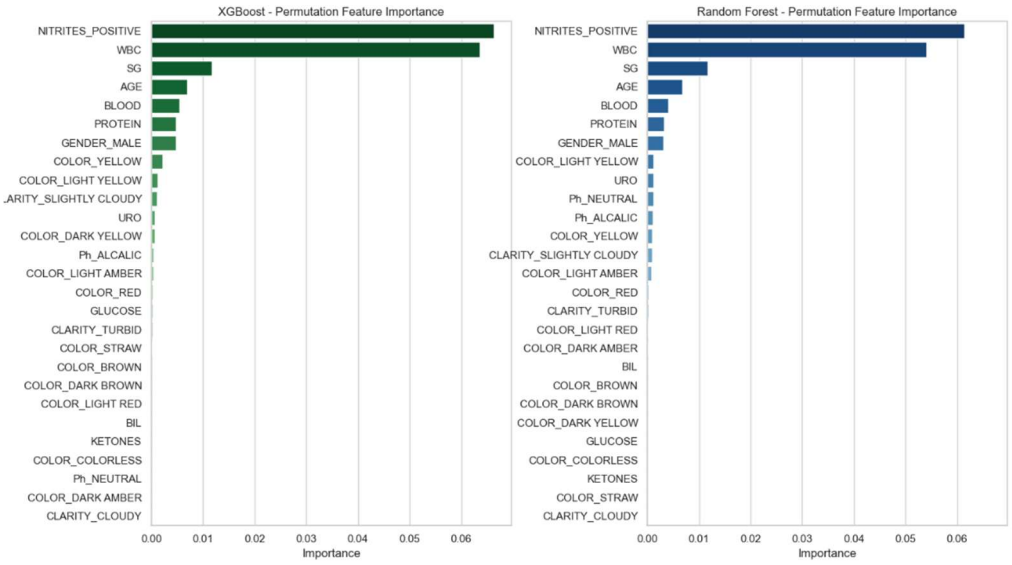

Figure 3. Permutation feature importance (PFI) for XGBoost (left) and Random Forest (right), based on accuracy loss across 10-fold cross-validation. Both models consistently ranked positive nitrites test (NITRITES_POSITIVE), white blood cell (WBC) count, and specific gravity (SG) among the most important features, suggesting strong predictive contribution of these urinalysis parameters.

Among the five evaluated ensemble models, XGBoost demonstrated the highest F1-score (0.648) and ROC AUC (0.887), as well as a favorable diagnostic odds ratio (DOR: 20.52). While Random Forest showed similar performance, XGBoost consistently achieved marginally superior results across multiple clinically relevant metrics, including balanced accuracy, specificity, and Matthews correlation coefficient. These findings, along with its capacity for fine-tuning through hyperparameter and threshold optimization, supported its selection for post-hoc enhancement.

To improve clinical sensitivity, the XGBoost model underwent threshold tuning using a probabilistic cutoff derived from the Youden index, following a similar approach to that described in previous machine learning studies for diagnostic model optimization (39–40).

Figure 4 shows the comparison of confusion matrices before and after threshold optimization. Applying the optimized threshold of 0.182 resulted in a marked reduction in false negatives (268 → 47), with a corresponding increase in true positives (1248 → 1469). This adjustment increased sensitivity from 70.2% to 87.9% and improved the negative predictive value (NPV) from 92.8% to 96.4%. However, this improvement came at the expense of increased false positives (532 → 1612), leading to a reduction in specificity (89.6% → 74.1%) and precision (60.9% → 44.1%). These changes reflect the expected trade-off between recall and precision, particularly relevant in screening contexts where missing infections may carry significant risk. ROC and precision–recall curves remained unchanged, as threshold tuning only alters classification decisions without affecting the predicted probabilities.

 doi:10.20944/preprints202504.0037.v1

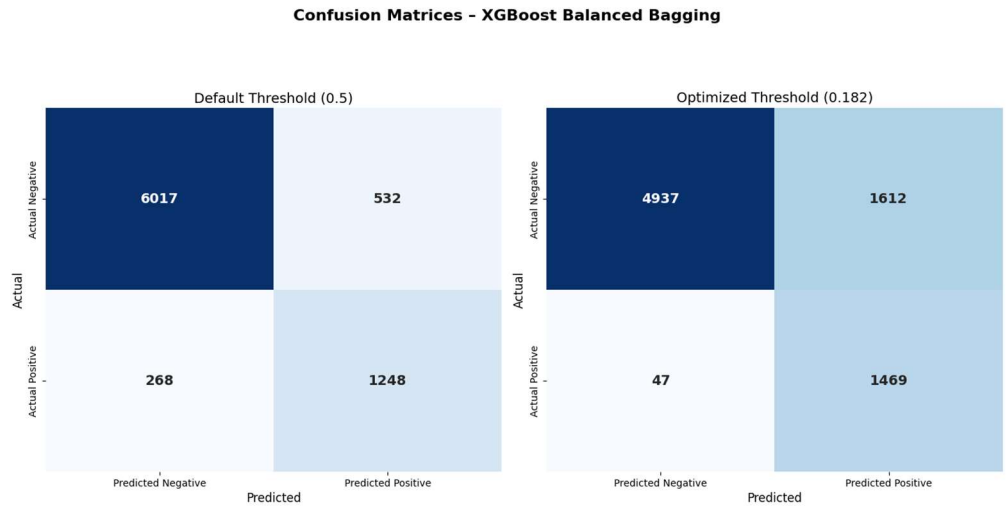

Figure 4. Confusion matrices for the XGBoost classifier before (left) and after (right) probability threshold optimization (cutoff = 0.182). The optimized threshold reduced false negatives while increasing true positives, sensitivity, and NPV, at the expense of lower specificity and precision.

The complete set of performance metrics before and after threshold optimization is summarized in Table 2. The results illustrate the expected trade-off introduced by post-hoc adjustment: sensitivity, negative predictive value, and diagnostic odds ratio were substantially improved, while specificity and precision declined. Notably, the ROC AUC remained unchanged, as threshold tuning does not alter predicted probabilities.

Table 2. Comparative performance of the XGBoost model before and after post-hoc threshold optimization (cutoff = 0.182).

Metric	XGBoost (Default Threshold)	XGBoost (Optimized Threshold)
Accuracy	0.859	0.767
Balanced Accuracy	0.799	0.810
Precision (PPV)	0.609	0.441
Recall (Sensitivity)	0.702	0.879
Specificity	0.896	0.741
Negative Predictive Value (NPV)	0.928	0.964
F1-score	0.652	0.587
ROC AUC	0.886	0.886
Matthews Corr. Coef. (MCC)	0.567	0.501
Positive Likelihood Ratio (PLR)	6.740	3.401
Negative Likelihood Ratio (NLR)	0.333	0.163
Diagnostic Odds Ratio (DOR)	20.250	20.893

4. Discussion

4.1. Dataset Description

The dataset used in this study was collected from the Laboratory Medicine Department of the General Hospital of Amfissa, Greece. It comprised urinalysis parameters, urine culture results, and demographic information from patients who underwent routine urinalysis and urine culture. Urinalysis data were obtained using the iChemVELOCITY Automated Urinalysis Analyzer (Leriva Diagnostics), an automated laboratory instrument employed in routine clinical practice. The analyzer

 doi:10.20944/preprints202504.0037.v1

evaluates urine dipstick tests using reflectance photometry and standardized colorimetric methods, providing operator-independent measurements of physical and chemical urine characteristics. The reported parameters included specific gravity (SG), pH, color, clarity, ketones, glucose, protein, nitrites, bilirubin, urobilinogen, white blood cells (WBC), and blood. All results were retrospectively extracted from the laboratory information system and fully de-identified prior to analysis. Urine culture results were classified as positive when bacterial growth exceeded 10^5 cfu/mL, and as non-positive otherwise, in accordance with established microbial threshold guidelines [16-18].

The dataset included 8,065 urine samples, of which 2,257 (28.0%) were classified as positive and 5,808 (72.0%) as non-positive. Of the total samples, 3,111 (38.6%) were from male and 4,954 (61.4%) from female patients. The mean patient age was 57.8 years (SD = 25.0). Among males, 22.6% of cultures were positive, while the positivity rate among females was 29.1%.

4.2. Data Preprocessing

Although none of the models achieved a negative likelihood ratio (NLR) below the conventional 0.1 threshold required for reliable clinical rule-out [41-42], most exhibited acceptable to high negative predictive values (NPV), particularly XGBoost (92.8%) and Random Forest (93.4%). These values suggest potential utility in supporting decisions to withhold empirical antibiotics in low-risk scenarios. However, for models with NPV below 90%, such as the Stacking and Voting classifiers (88.7% and 88.9%, respectively), clinical caution is warranted.

The limited sensitivity observed—especially in the high-specificity classifiers—is likely attributable not to algorithmic limitations but to the restricted clinical context available in the dataset. In the absence of symptomatology (e.g., fever, dysuria, or urgency), the models rely exclusively on urinalysis and demographic data, which may not capture the full clinical spectrum of UTI presentations. This highlights the importance of incorporating richer clinical features when higher recall is desired.

Previous research by Ourani et al. showed that using predefined reflex-to-culture criteria based on urinalysis could safely reduce unnecessary antibiotic prescriptions, supporting the idea that properly tuned ML tools may complement such clinical pathways [14].

XGBoost achieved the highest F1-score and ROC AUC among all models, indicating a favorable balance between case detection and diagnostic precision. Random Forest performed comparably. While Extra Trees also demonstrated strong specificity, its lower sensitivity resulted in a less favorable diagnostic trade-off.

4.3. Threshold Optimization

While XGBoost demonstrated the most balanced performance among the evaluated classifiers, its initial sensitivity (70.2%) highlights the need for further optimization in screening contexts. To address this, post-hoc threshold optimization was applied using the Youden index. Adjusting the classification threshold to 0.182—without retraining the model—led to a substantial increase in sensitivity (87.9%) and negative predictive value (96.4%), accompanied by a marked reduction in false negatives (from 268 to 47).

Although this adjustment resulted in decreased specificity (from 89.6% to 74.1%) and precision (from 60.9% to 44.1%), the overall trade-off was clinically favorable for screening scenarios where missing infections may carry significant consequences. Importantly, the ROC AUC remained unchanged (0.886), confirming that model discrimination was preserved and that the improvement was attributable solely to threshold adjustment. These findings support the practical utility of threshold tuning in real-world clinical applications, particularly in settings where model retraining is impractical or restricted. Such post-hoc optimization steps may be especially valuable in multidisciplinary workflows involving collaboration between clinicians, informatics specialists, and backend data analysts—bridging the gap between model development and operational deployment.

 doi:10.20944/preprints202504.0037.v1

4.4. Model Interpretability and Clinical Integration

Permutation feature importance (PFI) analysis revealed consistent key predictors across models, notably the positive nitrites test, WBC count, and SG. These features align with established clinical markers of UTI and enhance the transparency of model decisions, supporting integration into decision support tools. These findings mirror those reported by Chambliss et al., who demonstrated strong correlation between specific chemical urinalysis elements and positive cultures, highlighting the potential of automated workflows in initial infection screening [41].

All models were trained solely on structured, non-microscopic laboratory parameters— such as color, clarity, WBC, nitrites, and pH—along with demographic variables including age and gender. This configuration facilitates real-time integration into laboratory information systems (LIS) and electronic health records (EHR), supporting clinical decision-making in primary care settings and emergency departments.

Although the Voting Classifier achieved excellent specificity and a high positive likelihood ratio, its limited sensitivity restricts its usefulness in screening applications, where reliably identifying true cases is a central concern. Nevertheless, it may serve a valuable role in antimicrobial stewardship strategies aimed at reducing unnecessary antibiotic prescriptions. In our dataset, deploying such a high-specificity model could reduce empirical antibiotic use from a baseline of approximately 45% to just 10–11% of patients, representing a substantial step toward more judicious prescribing.

Beyond its diagnostic improvements, the threshold-optimized XGBoost model offers substantial advantages in clinical antimicrobial stewardship. In typical empirical practice, up to 45% of patients with nonspecific urinalysis findings may receive antibiotics without confirmed infection, leading to overtreatment and potential antimicrobial resistance. In contrast, our model recommended treatment for only 38.2% of cases, and just 19.99% of the total population would have received unnecessary antibiotics (i.e., false positives). This corresponds to an absolute reduction of 25 percentage points and a relative reduction of approximately 56% in overtreatment compared to empirical prescribing.

In clinical practice, treatment decisions for suspected UTI must almost always be made prior to the availability of culture results, as standard urine culture remains a time-consuming process. Although rapid molecular platforms (e.g., PCR-based diagnostics) are emerging, their real-world application remains in its early stages. Significant time and clinical experience will be required to evaluate their performance and determine whether they can be effectively adapted to meet the diagnostic needs of frontline physicians [43-47]. Additionally, their current limitations in availability, high cost, and limited clinical validation further restrict their widespread adoption.

In this context, machine learning–based decision support systems offer a practical and scalable alternative for improving diagnostic accuracy and guiding antibiotic prescribing—particularly in care environments where empirical decisions must be made before definitive culture results become available.

4.5. Limitations

This study is subject to several limitations. First, it was conducted using retrospective data from a single institution, which may limit generalizability across diverse populations or care settings. Second, the models relied exclusively on laboratory and demographic data, without incorporating clinical presentation, comorbidities, symptom duration, or prior antibiotic use—factors that can influence infection likelihood and diagnostic interpretation.

Although stratified 10-fold cross-validation was applied to minimize overfitting, external validation on independent datasets is essential before considering clinical deployment. Additionally, moderate class imbalance and the exclusion of incomplete records may have introduced bias or reduced model robustness.

Despite these limitations, this study presents a reproducible and interpretable machine learning framework using real-world urinalysis data. The results provide a foundation for future development of decision support tools and highlight the potential value of integrating predictive modeling into clinical laboratory workflows.

 doi:10.20944/preprints202504.0037.v1

4.6. Future Directions

Looking ahead, future research should focus on incorporating structured clinical features—such as urinary symptoms, comorbidities, and prior infection history—to enhance sensitivity and better capture the full clinical spectrum of UTI presentations. Prospective validation across diverse healthcare environments, including primary care and emergency departments, is essential to assess the generalizability and real-world impact of these models.

Moreover, the development of hybrid diagnostic frameworks that combine ML predictions with structured clinical criteria (e.g., dysuria, fever) or symptom-based rules may offer improved interpretability and greater clinical acceptance. Investigating the cost-effectiveness of such tools and their influence on prescribing behavior and antimicrobial resistance trends is a critical step toward sustainable clinical deployment. Finally, advances in explainable artificial intelligence (XAI) and user-centered interface design will be crucial for fostering trust and facilitating integration into laboratory information systems and electronic health records.

It is essential to emphasize that such tools are intended solely to complement—not replace—clinical judgment. The treating physician remains solely responsible for diagnosis and patient management, and these models should only be applied within clearly defined clinical pathways that preserve the primacy of medical decision-making.

5. Conclusions

This study demonstrates that machine learning models based on routine urinalysis and demographic data can support clinical decision-making in suspected urinary tract infections. Among the evaluated classifiers, XGBoost and Random Forest exhibited the most balanced diagnostic profiles, combining high specificity with interpretability and overall discriminative capacity.

The Voting Classifier achieved the highest specificity and positive likelihood ratio, suggesting potential utility in reducing unnecessary antibiotic prescriptions. However, the limited sensitivity observed across all models restricts their use as independent diagnostic tools, particularly in clinical settings where failure to detect infections may have consequences.

Given the absence of symptom-based clinical features in the dataset, we opted for post-hoc threshold optimization rather than retraining, in order to realign the XGBoost model with practical clinical needs. This adjustment achieved a more favorable balance between sensitivity and specificity, reducing false negatives while also lowering overtreatment rates compared to empirical prescribing.

These models—particularly XGBoost—may assist in identifying low-risk patients and guiding empirical treatment decisions in resource-constrained or high-volume settings. Their strength lies in stratifying patients according to infection risk using readily available structured data, thereby supporting antimicrobial stewardship and advancing individualized care.

Future research should focus on external validation in diverse populations, incorporation of clinical variables such as symptoms and comorbidities, and implementation studies assessing real-world impact. It is essential to maintain clear boundaries between algorithmic support and clinical responsibility; these tools are designed to inform—but not substitute—the diagnostic judgment and accountability of licensed healthcare professionals.

Author Contributions: Each author has contributed significantly to the study, as follows: A. S.: Conceptualization, Methodology, Data Collection, Writing—Original Draft. D. R.: Data Analysis, Data Preprocessing, Formal Analysis—Review & Editing. V. Z.s: Supervision. D. K.: Supervision—Review & Editing.

Funding: This research received no external funding

Institutional Review Board Statement: "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the General Hospital of Amfissa (116/ΔΣ-05.02.2025).

 doi:10.20944/preprints202504.0037.v1

Informed Consent Statement: "Patient consent was waived due to the retrospective nature of the study and the use of fully anonymized data, which did not allow for the identification of individual participants."

Data Availability Statement: The data presented in this study are not publicly available due to privacy and ethical restrictions, as they contain sensitive patient information collected from a clinical laboratory setting. Access to the dataset is restricted by the institutional policies of the General Hospital of Amfissa and cannot be shared publicly.

Conflicts of Interest "The authors declare no conflicts of interest."

Abbreviations

The following abbreviations are used in this manuscript:

UTIs	Urinary Tract Infections
ML	Machine Learning
SG	Specific Gravity
WBC	White Blood Cell
SMOTE	Synthetic Minority Over-sampling Technique
ADASYN	Adaptive Synthetic Sampling
XGBOOST	eXtreme Gradient Boosting
PPV	Positive Predictive Value
NPV	Negative Predictive Value
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
MCC	Matthews Correlation Coefficient
PLR	Positive Likelihood Ratio
NLR	Negative Likelihood Ratio
DOR	Diagnostic Odds Ratio
PFI	Permutation Feature Importance
LIS	Laboratory Information Systems
EHR	Electronic Health Records
PCR	Polymerase Chain Reaction
XAI	Explainable Artificial Intelligence

References

1. Sánchez, X.; Latacunga, A.; Cárdenas, I.; Jimbo-Sotomayor, R.; Escalante, S. Antibiotic Prescription Patterns in Patients with Suspected Urinary Tract Infections in Ecuador. PLoS ONE 2023, 18, e0295247. https://doi.org/10.1371/journal.pone.0295247
2. Fésüs, A.; Matuz, M.; Papfalvi, E.; Hambalek, H.; Ruzsa, R.; Tánczos, B.; Bácskay, I.; Lekli, I.; Illés, Á.; Benkő, R. Evaluation of the Diagnosis and Antibiotic Prescription Pattern in Patients Hospitalized with Urinary Tract Infections: Single-Center Study from a University-Affiliated Hospital. Antibiotics 2023, 12, 1689. https://doi.org/10.3390/antibiotics12121689
3. Medina, M.; Castillo-Pino, E. An Introduction to the Epidemiology and Burden of Urinary Tract Infections. Ther. Adv. Urol. 2019, 11, 1756287219832172. https://doi.org/10.1177/1756287219832172
4. Rowe, T.A.; Juthani-Mehta, M. Urinary Tract Infection in Older Adults. Aging Health 2013, 9, 515–528. https://doi.org/10.2217/ahe.13.38
5. Kolodziej, L.M.; Kuil, S.D.; de Jong, M.D.; Schneeberger, C. Resident-Related Factors Influencing Antibiotic Treatment Decisions for Urinary Tract Infections in Dutch Nursing Homes. Antibiotics 2022, 11, 140. https://doi.org/10.3390/antibiotics11020140
6. Centers for Disease Control and Prevention (CDC). Outpatient Antibiotic Prescribing in the United States. Available online: https://www.cdc.gov/antibiotic-use/hcp/data-research/antibiotic-prescribing.html (accessed on 28 March 2025)
7. Parente, D.; Shanks, D.; Yedlinksy, N.; Hake, J.; Dhanda, G. Machine Learning Prediction of Urine Cultures in Primary Care. Ann. Fam. Med. 2023, 21 (Suppl. 1), 4141. https://doi.org/10.1370/afm.21.s1.4141

 doi:10.20944/preprints202504.0037.v1

8. Dedeene, L.; Van Elslande, J.; Dewitte, J.; Martens, G.; De Laere, E.; De Jaeger, P.; De Smet, D. An Artificial Intelligence-Driven Support Tool for Prediction of Urine Culture Test Results. Clin. Chim. Acta 2024, 562, 119854. https://doi.org/10.1016/j.cca.2024.119854
9. Seheult, J.N.; Stram, M.N.; Contis, L.; Pontzer, R.E.; Hardy, S.; Wertz, W.; Baxter, C.M.; Ondras, M.; Kip, P.L.; Snyder, G.M.; Pasculle, A.W. 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, e0029123. https://doi.org/10.1128/jcm.00291-23
10. Jeng, S.L.; Huang, Z.J.; Yang, D.C.; et al. Machine Learning to Predict the Development of Recurrent Urinary Tract Infection Related to Single Uropathogen, Escherichia coli. Sci. Rep. 2022, 12, 17216. https://doi.org/10.1038/s41598-022-18920-3
11. Yen, C.C.; Ma, C.Y.; Tsai, Y.C. Interpretable Machine Learning Models for Predicting Critical Outcomes in Patients with Suspected Urinary Tract Infection with Positive Urine Culture. Diagnostics 2024, 14, 1974. https://doi.org/10.3390/diagnostics14171974
12. Choi, M.H.; Kim, D.; Park, Y.; Jeong, S.H. Development and Validation of Artificial Intelligence Models to Predict Urinary Tract Infections and Secondary Bloodstream Infections in Adult Patients. J. Infect. Public Health 2024, 17, 10–17. https://doi.org/10.1016/j.jiph.2023.10.021
13. Singh, P.; Taylor, A.; Kumar, R.; et al. Predicting Positive Urine Culture Results Using Machine Learning Algorithms Trained on Routine Urinalysis Data: A Multi-Center Retrospective Study. medRxiv 2024, preprint. https://doi.org/10.1101/2024.05.28.24306956
14. Ourani, M.; Honda, N.S.; MacDonald, W.; Roberts, J. Evaluation of Evidence-Based Urinalysis Reflex to Culture Criteria: Impact on Reducing Antimicrobial Usage. Int. J. Infect. Dis. 2021, 102, 40–44. https://doi.org/10.1016/j.ijid.2020.09.1471
15. Tomlinson, E.; Ward, M.; Cooper, C.; James, R.; Stokes, C.; Begum, S.; Watson, J.; Hay, A.D.; Jones, H.E.; Thom, H.; Whiting, P. Point-of-Care Tests for Urinary Tract Infections to Reduce Antimicrobial Resistance: A Systematic Review and Conceptual Economic Model. Health Technol. Assess. 2024, 28, 1–109. https://doi.org/10.3310/PTMV852
16. Sinawe, H.; Casadesus, D. Urine Culture. In: StatPearls [Internet]; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www.ncbi.nlm.nih.gov/books/NBK557569/ (accessed on 28 March 2025).
17. Bono, M.J.; Leslie, S.W.; Reygaert, W.C. Uncomplicated Urinary Tract Infections. In: StatPearls [Internet]; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www.ncbi.nlm.nih.gov/books/NBK470195/ (accessed on 28 March 2025).
18. Sabih, A.; Leslie, S.W. Complicated Urinary Tract Infections. In: StatPearls [Internet]; StatPearls Publishing: Treasure Island, FL, USA, 2025. Available online: https://www.ncbi.nlm.nih.gov/books/NBK436013/ (accessed on 28 March 2025).
19. Qiu, S.; Liu, Q.; Zhou, S.; Huang, W. Adversarial Attack and Defense Technologies in Natural Language Processing: A Survey. Neurocomputing 2022, 492, 278–307. https://doi.org/10.1016/j.neucom.2022.04.020
20. Roshan, S.E.; Asadi, S. Improvement of Bagging Performance for Classification of Imbalanced Datasets Using Evolutionary Multi-Objective Optimization. Eng. Appl. Artif. Intell. 2020, 87, 103319. https://doi.org/10.1016/j.engappai.2019.103319
21. Edward, J.; Rosli, M.M.; Seman, A. A Comprehensive Analysis of a Framework for Rebalancing Imbalanced Medical Data Using an Ensemble-Based Classifier. Pertanika J. Sci. Technol. 2024, 32, 6. https://doi.org/10.47836/pjst.32.6.12
22. Bozcuk, H.Ş.; Yıldız, M. A Balanced Bagging Classifier Machine Learning Model-Based Web Application to Predict Risk of Febrile Neutropenia in Cancer Patients. J. Cancer Sci. Clin. Ther. 2024, 8, 327–334. https://doi.org/10.26502/jcsct.5079256
23. Cearns, M.; Hahn, T.; Baune, B.T. Recommendations and Future Directions for Supervised Machine Learning in Psychiatry. Transl. Psychiatry 2019, 9, 271. https://doi.org/10.1038/s41398-019-0607-2
24. Mahesh, T.R.; Kumar, V.V.; Kumar, V.D.; Geman, O.; Margala, M.; Guduri, M. The Stratified K-Folds Cross-Validation and Class-Balancing Methods with High-Performance Ensemble Classifiers for Breast Cancer Classification. Healthc. Anal. 2023, 4, 100247. https://doi.org/10.1016/j.health.2023.100247

 doi:10.20944/preprints202504.0037.v1

25. Bey, R.; Goussault, R.; Grolleau, F.; Benchoufi, M.; Porcher, R. Fold-Stratified Cross-Validation for Unbiased and Privacy-Preserving Federated Learning. J. Am. Med. Inform. Assoc. 2020, 27, 1244–1251. https://doi.org/10.1093/jamia/ocaa096
26. Bradshaw, T.J.; Huemann, Z.; Hu, J.; Rahmim, A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. Radiol. Artif. Intell. 2023, 5, e220232. https://doi.org/10.1148/ryai.220232
27. Ghazwani, M.; Begum, M.Y. Computational Intelligence Modeling of Hyoscine Drug Solubility and Solvent Density in Supercritical Processing: Gradient Boosting, Extra Trees, and Random Forest Models. Sci. Rep. 2023, 13, 10046. https://doi.org/10.1038/s41598-023-37232-8
28. Jijo BT, Abdulazeez AM. Classification Based on Decision Tree Algorithm for Machine Learning. J Appl Sci Technol Trends. 2021;2(1):20–28. https://doi.org/10.38094/jastt20165
29. Mushtaq, Z.; Ramzan, M.F.; Ali, S.; Baseer, S.; Samad, A.; Husnain, M. Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. Mob. Inf. Syst. 2022, 2022, 6521532. https://doi.org/10.1155/2022/6521532
30. Srinivas, A.; Mosiganti, J.P. A Brain Stroke Detection Model Using Soft Voting-Based Ensemble Machine Learning Classifier. Meas. Sens. 2023, 29, 100871. https://doi.org/10.1016/j.measen.2023.100871
31. Ghasemieh, A.; Lloyed, A.; Bahrami, P.; Vajar, P.; Kashef, R. A Novel Machine Learning Model with Stacking Ensemble Learner for Predicting Emergency Readmission of Heart-Disease Patients. Decis. Anal. J. 2023, 7, 100242. https://doi.org/10.1016/j.dajour.2023.100242
32. Mohapatra, S.; Maneesha, S.; Mohanty, S.; Patra, P.K.; Bhoi, S.K.; Sahoo, K.S.; Gandomi, A.H. A Stacking Classifiers Model for Detecting Heart Irregularities and Predicting Cardiovascular Disease. Healthc. Anal. 2023, 3, 100133. https://doi.org/10.1016/j.health.2022.100133
33. Kent, P.; Hancock, M.J. Interpretation of Dichotomous Outcomes: Sensitivity, Specificity, Likelihood Ratios, and Pre-Test and Post-Test Probability. J. Physiother. 2016, 62, 231–233. https://doi.org/10.1016/j.jphys.2016.08.008
34. Marill, K.A. Diagnostic and Prognostic Test Assessment in Emergency Medicine: Likelihood and Diagnostic Odds Ratios. Emerg. Med. J. 2022, 39, 635–642. https://doi.org/10.1136/emermed-2022-212396
35. Huang, Y.; Yin, J.; Samawi, H. Methods Improving the Estimate of Diagnostic Odds Ratio. Commun. Stat. Simul. Comput. 2018, 47, 353–366. https://doi.org/10.1080/03610918.2016.1157183
36. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. Bioinformatics 2010, 26, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134
37. Molnar, C. Feature Importance. In: Interpretable Machine Learning [Internet]; 2019. Available online: https://christophm.github.io/interpretable-ml-book/feature-importance.html (accessed on 28 March 2025).
38. Schisterman, E.F.; Faraggi, D.; Reiser, B.; Hu, J. Youden Index and the Optimal Threshold for Markers with Mass at Zero. Stat. Med. 2008, 27, 297–315. https://doi.org/10.1002/sim.2993.
39. Tu, J.B.; Liao, W.J.; Liu, W.C.; Gao, X.H. Using Machine Learning Techniques to Predict the Risk of Osteoporosis Based on Nationwide Chronic Disease Data. Sci. Rep. 2024, 14, 5245.
40. Wang, J.; Wang, G.; Wang, Y.; Wang, Y. Development and Evaluation of a Model for Predicting the Risk of Healthcare-Associated Infections in Patients Admitted to Intensive Care Units. Front. Public Health 2024, 12, 1444176. https://doi.org/10.3389/fpubh.2024.1444176
41. Deeks, J.J.; Altman, D.G. Diagnostic tests 4: likelihood ratios. BMJ 2004, 329, 168. https://doi.org/10.1136/bmj.329.7458.168.
42. Florkowski, C.M. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin. Biochem. Rev. 2008, 29(Suppl 1), S83–S87. PMID: 18852864; PMCID: PMC2556590.
43. Zering, J.; Stohs, E.J. Urine polymerase chain reaction tests: stewardship helper or hinderance? Antimicrob. Steward. Healthc. Epidemiol. 2024, 4(1), e77. https://doi.org/10.1017/ash.2024.71. PMID: 38721490; PMCID: PMC11077600.
44. Kelly, B.N. UTI detection by PCR: Improving patient outcomes. J. Mass Spectrom. Adv. Clin. Lab. 2023, 28, 60–62. https://doi.org/10.1016/j.jmsacl.2023.02.006. PMID: 36895940; PMCID: PMC9988651.
45. Elia, J.; Hafron, J.; Holton, M.; Ervin, C.; Hollander, M.B.; Kapoor, D.A. The Impact of Polymerase Chain Reaction Urine Testing on Clinical Decision-Making in the Management of Complex Urinary Tract

 doi:10.20944/preprints202504.0037.v1

Infections. Int. J. Mol. Sci. 2024, 25(12), 6616. https://doi.org/10.3390/ijms25126616. PMID: 38928323; PMCID: PMC11203880.

46. Kapoor, D.A.; Holton, M.R.; Hafron, J.; Aljundi, R.; Zwaans, B.; Hollander, M. Comparison of Polymerase Chain Reaction and Urine Culture in the Evaluation of Patients with Complex Urinary Tract Infections. Biology 2024, 13(4), 257. https://doi.org/10.3390/biology13040257. PMID: 38666869; PMCID: PMC11048588.
47. Hao, X.; Cognetti, M.; Patel, C.; Jean-Charles, N.; Tumati, A.; Burch-Smith, R.; Holton, M.; Kapoor, D.A. The Essential Role of PCR and PCR Panel Size in Comparison with Urine Culture in Identification of Polymicrobial and Fastidious Organisms in Patients with Complicated Urinary Tract Infections. Int. J. Mol. Sci. 2023, 24(18), 14269. https://doi.org/10.3390/ijms241814269. PMID: 37762570; PMCID: PMC10531650.