



Published in final edited form as:

J Hosp Infect. 2025 August ; 162: 263–271. doi:10.1016/j.jhin.2025.04.024.

Comparative Analysis of Machine Learning Models for Predicting Hospital- and Community-Associated Urinary Tract Infections Using Demographic, Hospital, and Socioeconomic Predictors

Arash Arjmand¹, Majid Bani-Yaghoub^{1,*}, Gary Sutkin², Kiel Corkran¹, Susanna Paschal³

¹Division of Computing, Analytics and Mathematics, School of Science and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA.

²Department of Biomedical and Health Informatics, School of Medicine, University of Missouri-Kansas City, Kansas City, MO 64108, USA

³University Health, Kansas City Hospital 2301 Holmes Street, Kansas City, MO 64108, USA

Abstract

Background: Urinary tract infections (UTI) are among the most common infections encountered in both community and healthcare settings. Differentiating between community-associated UTI (CA-UTI) and healthcare-associated UTI (HA-UTI) is crucial for understanding their epidemiology, identifying risk factors, and developing appropriate treatment strategies. Machine learning (ML) techniques have shown significant potential in improving the accuracy of predicting these infections, enabling more effective interventions and better patient outcomes. While previous studies have demonstrated the utility of ML models in various healthcare settings, there is still a need for a comparative analysis of different ML approaches, particularly in distinguishing between CA-UTI and HA-UTI and assessing the risk of UTI among hospitalized patients.

Objective: Using 2019–2023 patient demographics, hospital, and socioeconomic data, this study aims to build, validate, and compare machine learning models—Decision Tree (DT), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) to differentiate between the incidences of HA-UTI and CA-UTI. Additionally, it seeks to identify key predictors of UTI using demographic, hospital, and socioeconomic variables.

Results: The **DT** model demonstrated the highest sensitivity, particularly in handling the highly imbalanced data of HAI, with a sensitivity of 87%. **LR** achieved the best overall accuracy, at

*Corresponding author baniyaghoubm@umkc.edu.

Model Implementation and Reproducibility

To replicate our study with a hypothetical dataset, researchers can download and run the source codes at the following GitHub link: https://github.com/AArjcode/UTI_MachineLearning_HAI_vs_CAI Detailed instructions on how to run the codes and how to interpret the outputs can be found in the supplementary word document entitled, “Instructions to Run Simulations”.

Ethics: This study was approved by the Institutional Review Board at the University of Missouri–Kansas City, IRB #2094337-KC.

Conflict of Interest Statement: None

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used Grammarly and ChatGPT to improve their writing style. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

95.9% for HA-UTI and 93.2% for HA-UTI vs. CA-UTI. **RF** performed best in cross-validation, reaching 99.1% for HA-UTI and 96.2% for HA-UTI vs. CA-UTI. **NN** showed the highest specificity, at 93.4%, for HA-UTI vs. CA-UTI. The **AUC** values further supported these findings, ranging from 71.9% for **NN** to 96% for **RF**, reflecting the robustness of these models across different annual datasets. Among patient demographics, hospital, and socioeconomic variables, all models consistently identified the nurse units (e.g., inpatient units and mental health units) as the most significant predictors of UTI. In addition to nurse units, **LR** and **DT** identified location (e.g., various clinics and medical centers) as a key predictor. For HA-UTI versus CA-UTI, variations were observed across the years, with patient age, median household income, and gender intermittently emerging as key predictors.

Conclusion: The predictive accuracy of the machine learning models is relatively the same, with some differences in sensitivity and specificity for identifying both HA-UTI vs. CA-UTI and HA-UTI. Nurse units consistently emerge as the most significant predictors across all years. The importance of all predictors, such as socioeconomic factors and location, varies from year to year, highlighting the need for incorporating those variables in the surveillance systems to optimize the accuracy of predictions.

Keywords

Urinary tract infection; community-associated infection; healthcare-associated infection; machine learning; socioeconomic factors

1. Introduction

Urinary tract infections (UTI) are among the most common infections encountered in both community and healthcare settings, contributing significantly to healthcare costs and patient morbidity. Differentiating between community-associated (CA-UTI) and healthcare-associated UTI (HA-UTI) is crucial for understanding their distinct epidemiological patterns and risk factors and developing effective treatment strategies.

The growing prevalence of multidrug-resistant organisms (MDROs) has made timely and accurate diagnosis of UTIs a crucial step to minimize the spread of MDROs. Traditional methods like urine dipstick tests and urine cultures are either too slow or lack diagnostic accuracy. New technologies, such as nucleic acid tests, mass spectrometry, biosensors, and microfluidic platforms, are improving the speed and accuracy of pathogen identification and antimicrobial susceptibility testing (AST). These emerging diagnostic technologies, which can deliver rapid results directly from urine samples, hold the potential to revolutionize UTI diagnosis and treatment by enabling precision medicine. With MDROs becoming a global health threat, accurate UTI prediction and early identification of resistance markers are essential for guiding antimicrobial therapy and improving patient outcomes [17].

Machine learning (ML) techniques have shown considerable promise in improving the accuracy of predicting various healthcare-associated infections, including UTI. Previous studies have demonstrated that ML models can predict hospital-onset UTI with greater Positive Predictive Value (PPV) than traditional methods, facilitating timely interventions and improving patient outcomes [1, 2]. When integrated with electronic health records

(EHRs), these models serve as powerful tools for identifying patients at risk of HA-UTI, enabling healthcare providers to implement preventive measures proactively [3, 9].

In both community and hospital settings, the microbial landscape and resistance patterns of UTI differ significantly, necessitating tailored approaches for diagnosis and treatment. Research has shown that HA-UTI often involves highly resistant pathogens compared to CA-UTI, highlighting the need for predictive models to distinguish between these two types [4, 8]. Furthermore, sociodemographic factors such as socioeconomic status and access to healthcare have been identified as significant predictors of UTI risk, particularly for multidrug-resistant infections, suggesting that these variables should be incorporated into predictive modeling [7].

While various ML models have been applied to predict UTI in specific contexts, such as recurrent UTI [10] or infections in intensive care units [5], there remains a gap in systematically comparing different ML techniques to optimize UTI prediction across diverse settings. This study aims to address this gap by training, validating, and comparing five supervised ML models: Decision Trees (DT), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). By evaluating these models, we aim to determine the most effective approach for distinguishing between CA-UTI and HA-UTI and to identify key predictors using a comprehensive set of demographic, hospital, and socioeconomic variables.

Through this research, we aim to contribute to more accurate UTI predictions, leading to better-targeted interventions and ultimately improved patient outcomes in both community and healthcare settings. By leveraging advanced ML techniques, we also hope to provide valuable insights that will enhance the management and prevention of UTI, particularly in the face of rising MDROs.

2. MATERIALS AND METHODS

2.1 Study Site and Patient Data.

We obtained daily patient data from 2019 to 2023 through Cerner Corporation, Kansas City. The data originates from University Health (UH) in Kansas City, Missouri, which encompasses 35 facilities, including two main campuses: University Health Truman Medical Center and University Health Lakewood Medical Center, as well as several primary care clinics. University Health Truman Medical Center is a 314-bed urban hospital with a Level I trauma center designation, offering both acute care and outpatient services. University Health Lakewood Medical Center, a 101-bed facility, provides similar comprehensive care. Both campuses are located in Kansas City, Missouri.

In this study, hospital location and nursing unit data were recoded based on the incidence and prevalence of infections to facilitate analysis. Locations with low incidence rates were categorized as “Other” to streamline the data into more manageable groups. Hospital locations were grouped into four main categories: UH Lakewood Medical Center, UH Truman Medical Center, UH Truman Medical Center BH, and Other. Similarly, nursing units were categorized according to their primary functions, which included Inpatient Units,

Mental Health Units, Surgery Units, Rehabilitation and Observation Units, Critical Care and Intensive Care Units, and Others. The Inpatient Units included wards such as Inpatient-3 Blue (indicated third floor of the hospital inpatient unit blue region), Inpatient-3 Gold, Inpatient-3 Red, Inpatient-4 Blue, Inpatient-4 Green, Inpatient-4 Red, and Inpatient-5 North. The Mental Health Units comprised Behavioral Health Services, Center of Mental Health, and Geriatric Psychiatry. Surgery Units included Surgery ICU, Med/Surg North, and Med/Surg South. Rehabilitation and Observation Units comprised Medical Rehab and 4 Gold Observation, while the Critical Care and Intensive Care Units included ICU, MICU II, Progressive Care Unit, and Cardiac Care Unit. All remaining nursing units were classified under the “Other” category.

This recoding process was undertaken to ensure that similar units were grouped together, thereby facilitating more meaningful comparisons across different hospital settings. Table 1 provides a summary of data across three categories: HA-UTI, CA-UTI, and UTI-free (i.e., patients without UTI) from 2019 to 2023.

2.2 Predicted Outcomes

We compared two primary groups: HA-UTI versus CA-UTI, and HA-UTI versus UTI-free. This approach allowed us to investigate differences in patient characteristics and outcomes between UTI associated with community versus healthcare settings, as well as the impact of UTI presence or absence in patients with prolonged hospital stays. It is widely accepted that a hospital-acquired infection (HAI) is defined by a minimum length of stay (LOS) of 48 hours. In this study, however, we adopted a slightly different approach. We categorized UTI as community-associated if the LOS was less than 96 hours (four days), and healthcare-associated if the LOS was four days or longer. This distinction is supported by research indicating that UTIs typically result in an additional four days (95% CI: 3.1–5.0 days) of hospitalization [37]. As such, hospitalized patients who develop a UTI are expected to remain in the hospital for at least four days on average. Due to the significant imbalance between UTI and non-UTI patients in the CA group, we did not compare UTI versus non-UTI patients within that group to avoid skewed results.

2.3 Model predictors

In this study, we included patient demographics, hospital variables, and socioeconomic factors. Specifically, demographic predictors included age, gender, ethnicity, marital status, and employment status. Hospital-related variables included the hospital location and the specific nurse unit. Socioeconomic factors, such as poverty level, the percentage of the population, level of education, median income, and population density, were not directly obtained from individual patients. Instead, these factors were assigned based on the ZIP code of the patient’s residence, with data acquired from [11]. This means that while we did not have individual-level socioeconomic data, we used the median household income and other factors associated with the patient’s ZIP code as proxies.

2.4 Supervised Machine Learning Models

We employed several other machine-learning techniques for predictive modeling, including Logistic Regression (LR), Decision Tree (DT), Neural Network (NN), Random Forest

(RF), and Extreme Gradient Boosting (XGBoost). We began with a multivariable logistic regression model [18], employing a stepwise approach to select the most relevant variables. This method iteratively adds variables to the model, ceasing when no further improvement is observed, as determined by the Akaike information criterion. DTs [19] are one of the most commonly used classification methods in machine learning, as they partition datasets into smaller, manageable groups based on specific criteria. This process of splitting is performed in a hierarchical structure, beginning from a root node and moving toward leaf nodes, where final classifications are made based on Boolean decisions at each step [12]. In our study, the DT segmented the patient population using features such as age, nurse unit, and socioeconomic factors, optimizing these splits at each decision point independently. The Gini coefficient was used to measure node impurity, allowing the model to choose the most effective splits. This approach is advantageous due to its simplicity and interpretability, especially in clinical settings where understanding the decision-making process is critical [12]. NNs, on the other hand, consist of multiple layers that transform input data through a combination of linear and nonlinear functions, capturing complex relationships between variables [20]. The NN used in our study was constructed with two hidden layers and employed a sigmoid activation function at the output layer. RF, an ensemble learning method, constructs multiple decision trees during training and aggregates their outputs by taking the mode of the classes (for classification) or the mean prediction (for regression [21]). This method combines decision trees with bootstrapping and aggregation techniques, providing a versatile framework capable of handling both regression and classification tasks. RF is particularly effective in improving prediction accuracy and reducing overfitting by averaging the results of individual trees. Additionally, due to its structure, RF is highly scalable and can be parallelized to handle large datasets, making it suitable for Big Data applications [13]. Additionally, XGBoost, a powerful gradient-boosting technique, improves predictive accuracy by sequentially adding new models to correct errors made by previous models [22]. Known for its scalability and efficiency, XGBoost has been optimized to handle large datasets with sparse or structured data and is renowned for its high computational performance by leveraging parallel and distributed computing. Its innovations, such as handling sparse data and implementing a weighted quantile sketch, make it highly effective for real-world machine-learning tasks [14].

To handle missing values, we applied an imputation technique as follows. We utilized the K-Nearest Neighbor (K-NN) imputation method to estimate and fill in missing values, following the approach described by Malarvizhi and Thanamani [15]. This step ensured that our dataset was complete and that the models could be trained effectively without bias due to missing data.

To address the class imbalance issue present in the data, we used the SMOTE (Synthetic Minority Over-sampling Technique [23]). Using the SMOTE, we enhanced the representation of the minority class, i.e., the incidences of HAIs, and minimized the likelihood of potentially biased model predictions. Briefly, the SMOTE generates synthetic samples by interpolating between randomly selected minority class examples and their nearest neighbors, identified using the K-NN algorithm [16]. This approach contrasts with traditional oversampling methods [24], which simply duplicate existing minority classes to reduce overfitting while improving model generalization.

After applying SMOTE, the dataset was split into training and testing sets, with 80% of the data allocated for training and the remaining 20% reserved for testing. Cross-validation was also conducted to refine the models further and ensure robustness in the evaluation. Model performance was assessed using the area under the curve (AUC) metric, and confusion matrices were generated for each model with a classification threshold set at 0.5. Additionally, the structure of the decision tree model was presented for transparency and interpretability, allowing for a clearer understanding of the decision-making process within the model.

3. Results

3.1 Descriptive statistics

Table 2 provides descriptive statistics of 2019–2023 patients with healthcare-associated infections (HAIs). We can observe distinct differences between those with and without urinary tract infections (UTI) based on various demographic, hospital, and socioeconomic factors. Patients with UTI tend to be older, with an average age of 63.5 years compared to 48.2 years for those without UTI. A higher percentage of UTI patients are female (65.7%) compared to those without UTI (56%). The length of stay (LOS) is significantly longer for UTI-free patients (78.5 days) compared to HA-UTI patients (20.1 days). This difference stems from the study's selection criteria, which included only patients hospitalized for four or more days in the UTI-free group, thereby excluding short-stay, non-UTI cases. Consequently, this group mainly comprises long-term hospitalized patients with chronic conditions or post-surgical complications. LOS also varied across years, with annual averages ranging from 55.1 to 91.3 days, contributing to the overall median of 78.5 days. UTI patients are associated with specific nurse units (58.7%) and hospital locations (72.9%) compared to those without UTI, who have much lower association with specific nurse units (8.3%) and different healthcare facilities (18.2%). In terms of socioeconomic factors, there are minor differences between the two groups.

Table S1 compares patients with healthcare-associated UTIs (HA-UTI) and community-associated UTIs (CA-UTI). There are notable differences based on demographic, hospital, and socioeconomic factors. HA-UTI patients are older compared to CA-UTI patients (average age of 63.5 years compared to 49.7 years). A higher percentage of CA-UTI patients are female (84.7%) compared to HA-UTI patients (65.7%). The length of stay is significantly longer for HA-UTI patients, averaging 20.1 days, compared to 0.6 days for CA-UTI patients. HA-UTI patients are more frequently located in the inpatient units (58.7%), mainly at UH Truman Medical Center (72.9%), compared to CA-UTI patients (4.0% and 30.1%, respectively). In terms of socioeconomic factors, both groups show similar levels of income, poverty, and education, with minor differences.

3.2 Model Accuracies

Tables 3 and S2 summarize the performance of the DT, NN, LR, RF, and XGBoost models across two predictive tasks: differentiating between HA-UTI and UTI-free cases (Table 3) and distinguishing between HA-UTI and CA-UTI (Table S2). The Logistic Regression (LR) model performs robustly in both tasks, with an overall accuracy of 95.9% (HAI with or

without UTI) and 93.2% (HA-UTI vs. CA-UTI). It achieves high specificity (96.4% in Table 3, 92.6% in Table S2) and AUC (94 and 96, respectively), positioning it as a reliable general-purpose model. The Decision Tree (DT) model excels in sensitivity, with 87% in Table 3 and 95% in Table S2, making it effective in identifying positive cases, though it trades off some Positive Predictive Value (PPV) (8% in Table 3, 56% in Table S2). The Neural Network (NN) model, while having lower sensitivity (48%–88%), compensates with strong specificity, particularly in Table S2 (93.4%).

The Random Forest (RF) model strikes a balance between sensitivity and specificity, with the PPV of 13.5% for the HA-UTI prediction (Table 3) and 60.2%–64.5% for HA-UTI vs. CA-UTI (Table S2). It also shows high cross-validation stability, achieving 99.1% (Table 3) and 96.2% (Table S2), demonstrating robustness across different data subsets. XGBoost performs well in both tasks, especially in AUC (92.1 in Table 3 and 95.7 in Table S2), highlighting its effectiveness for structured data and consistency in producing accurate predictions.

3.3 Logistics regression

We utilized stepwise logistic regression to develop predictive models in two distinct analyses: first, to differentiate between (HAI) and (CAI) among UTI patients, and second, to distinguish between (HAI) in patients with and without UTI. Stepwise logistic regression works by iteratively adding or removing variables based on their statistical significance in predicting the outcome. It begins with the most significant variable and adds one more variable at each step, testing the model's performance and removing any variable that no longer contributes meaningfully. This process continues until the best-fitting model, with only the most relevant predictors, is constructed.

The results of the stepwise logistic regression analysis of HA-UTI versus UTI-free are shown in Table 4. An Odds Ratio (OR) greater than 1 suggests an increased likelihood of HA-UTI, while an OR less than 1 indicates a decreased likelihood of HA-UTI. For instance, in 2022, patients in UH Truman Medical Center (UHTMC) had 12.66 times higher odds of having an HA-UTI compared to those not in other facilities, highlighting the elevated risk in these settings due to factors such as prolonged stays or invasive procedures. Also, in 2019, the Mental Health nurse unit showed a decreased likelihood of patients acquiring an HA-UTI compared to those without (OR = 0.38). This trend continues in 2022 and 2023, with ORs of 0.38 and 0.27, respectively. Age is also a significant predictor, with increasing odds of HA-UTI in older patients, with ORs of 1.65 in 2020 and 1.51 in 2022. Gender also plays a role, with females showing a lower likelihood of acquiring HA-UTI compared to males without UTI, as seen in 2021 (OR = 0.33) and 2022 (OR = 0.41). Both UH Truman Medical Center (UHTMC) and UH Lakewood Medical Center (UHLMC) show higher odds of HAI with UTI in 2020, with ORs of 12.66 and 7.81, respectively, indicating a greater risk of acquiring HA-UTI in these locations. Ethnicity emerges as a predictor only in 2022, with an OR of 3.18, suggesting that certain ethnic groups may face a higher risk of acquiring UTI among HAI patients.

Similarly, Table S3 identifies significant predictors for HA-UTI versus CA-UTI. An Odds Ratio (OR) greater than 1 suggests an increased likelihood of HA-UTI, while an OR

less than 1 indicates a decreased likelihood of HA-UTI. For instance, in 2019, patients in mental health units had 45 times higher odds of having an HA-UTI versus CA-UTI compared to those not in these units (OR = 45.32), highlighting the elevated risk in these settings due to factors such as prolonged stays or invasive procedures. The inpatient nurse unit also shows a positive association with HA-UTI, especially in 2022 (OR = 1.83), suggesting that patients in these units face higher infection risk, potentially due to longer hospital stays or more complex care needs. Both UH Lakewood and UH Truman Medical Centers show varying degrees of positive associations with HA-UTI. For example, UH Truman demonstrated strong associations with HA-UTI in 2020 (OR = 17.63), indicating an increased risk in these hospital locations. Age consistently appears as a predictor, with increasing age associated with a slightly higher likelihood of HA-UTI (OR ~1.4 in 2020 and 2021). Overall, this analysis underscores the importance of specific nurse units and hospital locations in influencing the risk of HA-UTI compared to CA-UTI, and it suggests the need for targeted interventions, particularly in mental health and inpatient units, as well as ongoing monitoring of risk factors like age.

3.4 Classification and Regression Tree

Normalized importance in CRT represents how much each variable contributes to improving the model's classification accuracy. Table 5 shows the normalized importance of variables for distinguishing between HA-UTI and CA-UTI from 2019 to 2023. The Nurse Unit (Other) consistently ranks as the most significant predictor across all years, with importance scores ranging from 73.3% in 2020 to 97.3% in 2021, underscoring its central role in distinguishing HA-UTI from CA-UTI. Other variables such as Age, Poverty, Level of education, and Median Income show varying degrees of importance over time but are notably less influential compared to the Nurse Unit (Other). For instance, Age reaches its highest importance in 2022 (2.9%), and Poverty is particularly significant in 2020 (11.3%). Facility locations, like UH Truman Medical Center (UHTMC), also feature intermittently, particularly in 2019 (0.9%) and 2020 (5.5%), reflecting its shifting dynamics in predicting HA-UTI vs. CA-UTI over the years. This variation underscores the evolving nature of the contributing factors, especially for socioeconomic variables, which are more prominent in some years than others. Overall, Table 5 clearly indicates that the Nurse Unit (Other) remains the dominant factor, with other demographic and socioeconomic variables playing supporting but less consistent roles over time.

Similarly, Table S4 presents the normalized importance factors for predicting HA-UTI versus UTI-free from 2019 to 2023. Again, the Nurse Unit (Other) consistently ranks as the most influential predictor across all years, with importance scores ranging from 58.2% in 2020 to 93.6% in 2021, underscoring its central role in distinguishing between HA-UTI and UTI-free cases. Other variables, such as Gender (Male), Age, and Income, contribute less consistently. For example, Income shows a strong influence in 2020 (22.7%), while Gender (Male) holds importance in 2019 (3.3%) and 2022 (3.4%). Age also plays a role, particularly in 2022 (6.7%). Hospital locations, such as UH Truman Medical Center (UHTMC) and UH Lakewood Medical Center (UHLMC), are relevant in some years, with UHTMC showing a 7.7% importance in 2020 and UHLMC contributing 3.0% in 2023. Socioeconomic factors,

such as Population Density and Poverty, emerge sporadically, with Population Density having a higher significance in 2022 (3.2%) and Poverty playing a role in 2023 (1.5%).

Overall, Table S4 highlights the dominant role of the Nurse Unit (Other), with demographic, socioeconomic, and hospital location factors contributing variably over the years.

The tree diagrams corresponding to each decision tree model are provided in Section S5 of the supplementary document. The size of each compartment reflects the proportion of the population that belongs to the respective group. Additionally, distinct colour codes have been applied to differentiate between HA-UTI, CA-UTI, and UTI-free cases. As illustrated in Figures S3–S12, the tree diagram results align with the normalized variable importance rankings shown in Tables 5 and S4.

The decision tree in figure 1 classifies UTIs as either Healthcare-Associated Infection (HAI) or Community-Acquired Infection (CAI) based on several factors. The root node splits based on Nurse Unit (Other). If the nursing unit is classified as “Other,” the left branch is taken. This branch further splits into “Rehabilitation and Observation Units,” then “Nurse Unit Surgery,” and finally “AGE.” The right branch, where the nursing unit is *not* “Other,” splits based on “Poverty,” then “Location: UH Lakewood Medical Center,” and “Graduate degree.” The tree suggests that the type of nursing unit is a strong initial predictor. Rehabilitation and Observation Units show a higher propensity for HAI. Within the “Other” nursing unit category, surgical units also lean towards HAI, though age plays a role too. For cases not in the “Other” category, poverty, location at UH Lakewood Medical Center, and lower levels of graduate education are associated with CAI. The colours blue and orange represent HAI and CAI, respectively. The Gini index indicates the impurity of each node, with higher values indicating a more even mix of HAI and CAI.

4. Discussion

This study evaluates the strengths and limitations of five commonly used machine learning models in health and medical research. Namely, we applied Decision Tree (DT), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) to distinguish between HA-UTI and CA-UTI as well as between HA-UTI and UTI-free, using patient demographic, hospital, and socioeconomic variables. Our findings show that while LR generally achieved the highest overall accuracy and AUC, the RF model offered a balanced performance with superior sensitivity, making it particularly valuable in detecting cases of HAI. However, the DT model was superior not only for its higher sensitivity but also for its interpretability, making it easier to understand and communicate the risk factors to healthcare providers.

DT’s higher sensitivity highlights its utility in scenarios where minimizing missed infections is critical, while **NN** demonstrated higher specificity, reducing the likelihood of false positives. This distinction underscores the need to balance sensitivity and specificity depending on the clinical context. For instance, in situations where the consequences of missed **HA-UTI** are severe, sensitivity may be prioritized, making **DT** a more suitable choice.

When considering PPV and F1 score, **LR** and **RF** performed better, suggesting they are more reliable in correctly identifying true positive cases and balancing PPV and sensitivity (see Table3). This makes them strong candidates for applications where both PPV and sensitivity are crucial. However, the interpretability of **DT**, which clearly explains the influence of each predictor, provides an important advantage for practical implementation in healthcare settings, where clear communication of risk factors is essential.

The choice of evaluation metrics is particularly important in imbalanced datasets, such as in this study, where the HA-UTI group is significantly smaller than the UTI-free group. Accuracy alone may not provide a meaningful assessment of model performance in such cases, as a model that predominantly predicts the majority class could achieve high accuracy without effectively identifying the minority class.

Thus, PPV and sensitivity are more appropriate measures in this scenario. Sensitivity evaluates the model's ability to correctly identify HA-UTI cases, while PPV ensures that a high proportion of predicted HA-UTI cases are truly positive. The observed discrepancy between accuracy and PPV across models (Table 3) highlights this issue—some models achieve high accuracy but low PPV, indicating they may not be as effective in clinical decision-making, where false positives can lead to unnecessary treatments or interventions. This reinforces the importance of balancing sensitivity, PPV, and accuracy when evaluating model performance for healthcare applications.

All machine learning models ranked Nurse Unit as a significant predictor of HA-UTI likely due to several factors. First, nursing units often serve populations with varying health conditions and comorbidities, which may predispose certain groups to a higher risk of developing a HA-UTI [25–27, 35]. For example, units that care for patients with compromised immune systems [28, 29], those undergoing invasive procedures [30], or individuals with limited mobility [31] may inherently face higher UTI incidence. Second, HA-UTI rates may reflect broader health conditions, as UTI can be a marker of overall health and functional status, both of which can differ across nursing units [25, 28, 30].

Furthermore, differences in catheter management practices [32, 33] among staff in different nursing units could also contribute to this variability. Catheter use protocols, hygiene practices, and frequency of catheter changes might differ between units, further influencing HA-UTI rates [32, 33]. Other factors include nurse-to-patient ratio [34], overcrowding, higher patient turnover [27], and decreased monitoring time [34, 36]. As shown in Table S3, Rehabilitation and Inpatient Units were identified as significant predictors of HA-UTI, reinforcing the role of specific hospital environments in infection risk. These factors together highlight the complex, multifactorial nature of UTI risk as it relates to Nurse Unit.

5. Limitations

This study has several limitations that should be acknowledged. First, variables associated with the Socioeconomic Status (SES) were based on the ZIP code of the patients rather than individual-level data, which may not accurately reflect the true SES of each patient. Second, we defined community-associated UTI as those occurring in patients with a hospital length

of stay of fewer than 4 days, and healthcare-associated UTI as those in patients with a stay of 4 days or more; however, this criterion is not definitive and may not capture all cases accurately. Third, we did not model the CA-UTI versus UTI-free patients due to extremely low incidences of CA-UTI, which would have affected the reliability of the models. Additionally, our dataset, particularly the HA-UTI versus UTI-free, was highly imbalanced. Although we employed techniques such as SMOTE to address this issue partially, imbalance remained a challenge throughout the study. Lastly, our analysis was limited to the variables available in the dataset, as we were constrained by the absence of certain potentially relevant factors that could have provided a more comprehensive understanding of the predictors and outcomes.

Another key limitation of this study is the absence of additional clinical data, such as past hospitalization history, antibiotic use, previous urinary tract infections, use of urinary catheters, kidney disease, diabetes, and neurological symptoms. Incorporating these variables into the models could have improved their accuracy and provide a more comprehensive understanding of the factors influencing patient outcomes. Nonetheless, such data was not available at the time of study.

Additionally, the higher percentage of UTI cases at UH Truman Medical Center may be attributed to the hospital's patient demographics, including higher acuity and comorbidities, as well as differences in UTI screening and diagnostic protocols compared to other participating hospitals. These factors likely contributed to a greater detection rate of UTIs at this facility.

6. Conclusion

The machine learning models demonstrated relatively high predictive accuracy. Among all models, the nurse unit consistently emerged as the most significant predictor, while the importance of socioeconomic factors and locations varied from year to year. These findings underscore the potential of machine learning models to improve infection prediction and patient outcomes, with the optimal model choice depending on specific clinical objectives, such as prioritizing sensitivity, specificity, or interpretability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment:

This study was supported by the Center for Disease Control and Prevention under grant number 1U01CK000671-01.

References

1. Zachariah P, Sanabria E, Liu J, Cohen B, Yao D and Larson E, 2020. Novel strategies for predicting healthcare-associated infections at admission: Implications for nursing care. *Nursing Research*, 69(5), pp.399–403. [PubMed: 32604154]

2. Taylor RA, Moore CL, Cheung KH and Brandt C, 2018. Predicting urinary tract infections in the emergency department with machine learning. *PLOS ONE*, 13(3), p.e0194085. [PubMed: 29513742]
3. Møller JK, Sørensen M and Hardahl C, 2021. Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: A retrospective cohort study. *PLOS ONE*, 16(3), p.e0248636. [PubMed: 33788888]
4. Medina-Polo J, Naber KG and Johansen TEB, 2021. Healthcare-associated urinary tract infections in urology. *International Journal of Urology*, 28(4), pp.359–366.
5. Barchitta M, Maugeri A, Favara G, Riela PM, Gallo G and Mura I, 2021. A machine learning approach to predict healthcare-associated infections at intensive care unit admission: Findings from the SPIN-UTI project. *Journal of Hospital Infection*, 112, pp.94–101.
6. Aguilar-Duran S, Horcajada JP, Sorli L, Montero M, Salvado M, Grau S, Gomez J and Knobel H, 2012. Community-onset healthcare-related urinary tract infections: Comparison with community and hospital-acquired urinary tract infections. *Journal of Infection*, 64(3), pp.258–264.
7. Casey JA, Rudolph KE, Robinson SC, Bruxvoort K, Raphael E, Hong V, Pressman A, Morello-Frosch RX and Tartof SY, 2020. Sociodemographic inequalities in urinary tract infection in 2 large California health systems. *Clinical Infectious Diseases*, 71(1), pp.e114–e121.
8. Mancini A, Pucciarelli S, Lombardi FE, Barocci S and Pauri P, 2019. Differences between community- and hospital-acquired urinary tract infections in a tertiary care hospital. *Journal of Global Antimicrobial Resistance*, 17, pp.64–68.
9. Park JI, Bliss DZ, Chi CL, Delaney CW and Westra BL, 2021. Knowledge discovery with machine learning for hospital-acquired catheter-associated urinary tract infections. *Nursing Research*, 70(2), pp.118–125.
10. Jeng SL, Huang ZJ, Yang DC, Teng CH and Wang MC, 2021. Machine learning to predict the development of recurrent urinary tract infection related to single uropathogen, *Escherichia coli*. *Scientific Reports*, 11(1), p.3452. [PubMed: 33568737]
11. United States Zip Codes <https://www.unitedstateszipcodes.org>.
12. Charbuty B and Abdulazeez A, 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), pp.20–28.
13. Genuer R, Poggi JM, Tuleau-Malot C and Villa-Vialaneix N, 2017. Random forests for big data. *Big Data Research*, 9, pp.28–46.
14. Chen T and Guestrin C, 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
15. Malarvizhi R and Thanamani AS, 2012. K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev*, 5(1), pp.5–7.
16. Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP, 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp.321–357.
17. Davenport M, Mach KE, Shortliffe LMD, Banaei N, Wang TH and Liao JC, 2017. New and developing diagnostic technologies for urinary tract infections. *Nature Reviews Urology*, 14(5), pp.296–310. [PubMed: 28248946]
18. Stoltzfus JC (2011). Logistic regression: a brief primer. *Academic emergency medicine*, 18(10), 1099–1104. [PubMed: 21996075]
19. Rana KK (2014). A survey on decision tree algorithm for classification. *International journal of Engineering development and research*, 2(1), 1–5.
20. Goldberg Y (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57, 345–420.
21. Biau G (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1), 1063–1095.
22. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, ... & Zhou T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4–2*, 1(4), 1–4.
23. Nitesh VC (2002). SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*, 16(1), 321.

24. Gosain A, & Sardana S (2017, September). Handling class imbalance problem using oversampling techniques: A review. In 2017 international conference on advances in computing, communications and informatics (ICACCI) (pp. 79–85). IEEE.
25. Briongos-Figuero LS, Gómez-Traveso T, Bachiller-Luque P, Domínguez-Gil González M, Gómez-Nieto A, Palacios-Martín T, ... & Pérez-Castrillón JL. (2012). Epidemiology, risk factors and comorbidity for urinary tract infections caused by extended-spectrum beta-lactamase (ESBL)-producing enterobacteria. *International journal of clinical practice*, 66(9), 891–896. [PubMed: 22897466]
26. Laudisio A, Marinosci F, Gemma A, Bartoli IR, Montenegro N, & Incalzi RA (2017). The burden of comorbidity is associated with antibiotic resistance among institutionalized elderly with urinary infection: a retrospective cohort study in a single Italian nursing home between 2009 and 2014. *Microbial Drug Resistance*, 23(4), 500–506. [PubMed: 27525808]
27. Sara SM, Thota RC, Uddin MYS, Bani-Yaghoub M, Sutkin G, & Abourraja MN (2024). Patient flow modeling and simulation to study HAI incidence in an Emergency Department. *Smart Health*, 32, 100467.
28. De Pinho AM, Lopes GS, Ramos-Filho CF, da R Santos O, De Oliveira MP, Halpern M, ... & Schechter M. (1994). Urinary tract infection in men with AIDS. *Sexually Transmitted Infections*, 70(1), 30–34.
29. Jarzembowski T, Daca A, & Dębska-Ślizień MA (Eds.). (2018). *Urinary Tract Infection: The Result of the Strength of the Pathogen, or the Weakness of the Host*. BoD—Books on Demand.
30. Goldberg B, & Jantusch B (2016). Urinary tract infection. In *Clinical pediatric nephrology* (pp. 983–1008). CRC Press.
31. Rogers MA, Fries BE, Kaufman SR, Mody L, McMahon LF, & Saint S (2008). Mobility and other predictors of hospitalization for urinary tract infection: a retrospective cohort study. *BMC geriatrics*, 8, 1–8. [PubMed: 18194534]
32. Cravens DD, & Zweig S (2000). Urinary catheter management. *American family physician*, 61(2), 369–376. [PubMed: 10670503]
33. Nicolle LE (2005). Catheter-related urinary tract infection. *Drugs & aging*, 22, 627–639. [PubMed: 16060714]
34. Zalamar M (2024). *The Impact of Nurse-to-Patient Ratio on Patient Outcomes and Quality of Care: A Healthcare Policy Analysis and Advocacy* (Doctoral dissertation, Jacksonville University).
35. Al Kadria HM, El-Metwally AA, Al Sudairy AA, Al-Dahash RA, Al Khateeb BF, & Al Johani SM (2024). Antimicrobial Resistance Among Pregnant Women with Urinary Tract Infections is on Rise: Findings from Meta-analysis of Observational Studies. *Journal of Infection and Public Health*.
36. Thota RC, Sara SM, Uddin MYS, Bani-Yaghoub M, & Sutkin G (2024, May). Accurate Estimation of Individual Transmission Rates Through Contact Analytics Using UWB Based Indoor Location Data. In 2024 International Conference on Smart Applications, Communications and Networking (SmartNets) (pp. 1–8). IEEE.
37. Mitchell BG, Ferguson JK, Anderson M, Sear J, & Barnett A (2016). Length of stay and mortality associated with healthcare-associated urinary tract infections: a multi-state model. *Journal of Hospital Infection*, 93(1), 92–99. [PubMed: 26944900]

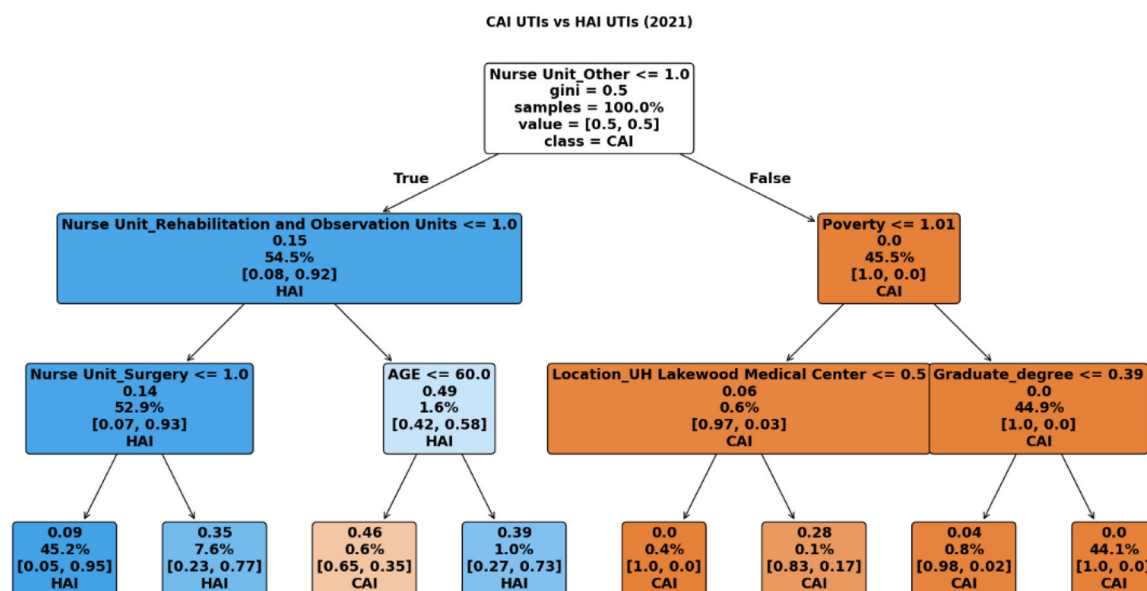


Figure 1:
Decision Tree for CAI UTI vs HAI UTI (2021)

Table1.

University Health patient flow according to HA-UTI, CA-UTI, and HA-free during 2019–2023

	2019	2020	2021	2022	2023
HA-UTI	383	399	410	357	328
CA-UTI	2,957	2,592	3,346	3,535	3,179
UTI-free	32,025	35,040	37,008	37,137	43,819

Table 2.

Descriptive statistics of 2019–2023 data organized according to patient demographics, hospital characteristics, and socioeconomic factors. Median values are reported for each group across the years, with minimum and maximum values presented in parentheses.

Model	Patients with HA-UTI	Patients without UTI	P-value
Population Size	383 (328, 410)	37008 (32025, 43819)	
Average Age ³	63.5 (58.9, 65.3)	48.2 (44.9, 51.3)	<0.01
Female ³ (%)	65.7 (61.6, 71.0)	56.0 (55.3, 59.1)	<0.01
Married (%)	16.4 (12.5, 20.7)	20.1 (18.2, 21.7)	<0.01
Employed (%)	99.7 (92.4, 100)	97.0 (84.3, 99.7)	Varies*
Hispanic (%)	8.5 (6.3, 10.1)	9.8 (9.0, 14.9)	Varies*
Length of Stay ⁵	20.1 (16.2, 27.0)	78.5 (55.1, 91.3)	<0.01
Nurse Unit ¹ (%)	58.7 (53.9, 63.7)	8.3 (5.9, 9.3)	<0.01
Location ² (%)	72.9 (62.7, 74.2)	18.2 (16.4, 19.5)	<0.01
Median Income	50,180.1 (49,141.7, 51,620.2)	49,371.1 (48,956.4, 50,236.3)	Varies*
Poverty (%)	18.7 (18.1, 19.3)	19.7 (18.8, 19.8)	Varies*
Graduate Degree (%)	7.9 (7.5, 8.2)	8.2 (8.0, 8.2)	Varies*
Population Density ⁴	2,870 (2,760, 2,936)	2,900 (2,735, 3,075)	Varies*

¹Percent of patients in the Inpatient Nurse Unit, this percentages reflect only the Inpatient Nurse Unit, with the remaining patients assigned to other units not explicitly shown in the table. ,

²Percent of patients in UH Truman Medical Center,

³Age, gender, marital status, length of stay, nurse unit and location were significant at the level of 0.01 for all 5 years 2019–2023. Varies” indicates that the statistical significance of these variables fluctuated across study years. For all five years (2019–2023), results for employment, ethnicity, median income, poverty, graduate degree, and population density were not consistently significant across all years.

⁴Number of individuals per square mile.

⁵This variable was excluded from all machine learning models. The unusually long length of stay for some patients was attributed to cases involving individuals in custody, where their hospital stays remained open beyond their discharge. The length of stay was capped at 365 days to address this issue in the analysis. This variable was not used in any of the machine learning modeling

Table 3.

Evaluation Results of Decision Tree (DT), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) for HA-UTI versus UTI-free cases.

Metrics	DT	NN	LR	RF	XGBoost
Accuracy	88.6 (87.4, 92.8)	95.7 (94.1, 97.3)	95.9 (95.4, 96.8)	95.8 (94.7, 97)	94.9 (93.1, 96.2)
Sensitivity	87 (74, 90)	48 (45, 55)	49 (45, 52)	58.1 (47.9, 62.3)	59 (49.3, 62.1)
Specificity	88.6 (87.4, 92.9)	96.2 (94.6, 97.7)	96.4 (96, 97.2)	96.3 (95, 97.4)	95.4 (93.5, 96.3)
PPV	8 (7, 9)	13 (9, 15)	13 (11, 15)	13.5 (11.9, 16.4)	11.8 (9, 12.6)
F1 Score	14.7 (12.9, 16.3)	21 (15.2, 22.7)	20 (18, 23.3)	25.6 (21.3, 35.1)	19.6 (15.6, 20.4)
CV ¹	94.5 (93.4, 96.4)	NA	NA	99.1 (98.7, 99.3)	99 (98.6, 99.2)
AUC ²	89.4 (87.4, 93)	71.9 (71.3, 75.6)	94 (93, 95)	89.2 (87.4, 90.3)	92.1 (91.1, 94)

¹Cross Validation (CV).

²Area Under the Curve (AUC) associated with the Receiver Operating Characteristic (ROC) curves.

Table 4.

Stepwise Logistic Regression Estimates for Predicting HA-UTI versus UTI-free across various years. The values are associated with different units and facility locations.

Predictors	2019	2020	2021	2022	2023
Other ¹	0.01	0.06	0.00	0.01	0.00
Mental Health ¹	0.38	— ³	—	0.38	0.27
Rehabilitation ¹	—	—	0.18	—	—
Surgery ¹	—	1.92	1.57	1.91	—
UHTMC ² (UHLMC ³)	—	12.66 (7.81)	—	—	—
UHTMC-BH	—	5.93	0.12	—	—
Age	—	1.65	1.41	1.51	1.28
Gender	0.43	0.46	0.33	0.41	0.44

¹ *Nurse Unit is divided into* Inpatient Units, Mental Health Units, Surgery Units, Rehabilitation and Observation Units, Critical Care and Intensive Care Units, and Others,

² University Health Truman Medical Center (UHTMC). Note that Mental Health is an outpatient unit,

³ “—” denotes that the variable was not selected or did not meet statistical significance for the given year in the stepwise logistic regression model

Table 5.

Normalized Importance of variables in prediction of HA-UTI vs CA-UTI across years 2019–2023

2019	2020	2021	2022	2023
Other ¹ (93.1)	Other ¹ (73.3)	Other ¹ (97.3)	Other ¹ (91.6)	Other ¹ (90.1)
Age (2.1)	Poverty (11.3)	Age (0.7)	Age (2.9)	Age (1.6)
UHTMC (0.9)	UHTMC (5.5)	Degree ³ (0.7)	Density ⁴ (2.8)	Ethnicity ⁶ (1.3)
Income ⁵ (0.8)	Age (3.1)	Density ⁴ (0.3)	Poverty (0.8)	E. Status ⁷ (1.2)
Degree ³ (0.7)	Income ⁵ (2.0)	Median Income (0.2)	Income ⁵ (0.5)	M. Status ⁸ (1.1)

¹ *Nurse Unit is divided into* Inpatient Units, Mental Health Units, Surgery Units, Rehabilitation and Observation Units, Critical Care and Intensive Care Units, and Others,

² University Health Truman Medical Center (UHTMC),

³ Percent Graduate Degree,

⁴ Population Density,

⁵ Median Household Income,

⁶ Hispanic vs. Non-Hispanic,

⁷ Employment Status,

⁸ Marital Status