

RESEARCH

Open Access



Prediction of urinary tract infection using machine learning methods: a study for finding the most-informative variables

Sajjad Farashi^{1,3*} and Hossein Emad Momtaz^{2,3*}

Abstract

Background Urinary tract infection (UTI) is a frequent health-threatening condition. Early reliable diagnosis of UTI helps to prevent misuse or overuse of antibiotics and hence prevent antibiotic resistance. The gold standard for UTI diagnosis is urine culture which is a time-consuming and also an error prone method. In this regard, complementary methods are demanded. In the recent decade, machine learning strategies that employ mathematical models on a dataset to extract the most informative hidden information are the center of interest for prediction and diagnosis purposes.

Method In this study, machine learning approaches were used for finding the important variables for a reliable prediction of UTI. Several types of machines including classical and deep learning models were used for this purpose.

Results Eighteen selected features from urine test, blood test, and demographic data were found as the most informative features. Factors extracted from urine such as WBC, nitrite, leukocyte, clarity, color, blood, bilirubin, urobilinogen, and factors extracted from blood test like mean platelet volume, lymphocyte, glucose, red blood cell distribution width, and potassium, and demographic data such as age, gender and previous use of antibiotics were the determinative factors for UTI prediction. An ensemble combination of XGBoost, decision tree, and light gradient boosting machines with a voting scheme obtained the highest accuracy for UTI prediction (AUC: 88.53 (0.25), accuracy: 85.64 (0.20)%), according to the selected features. Furthermore, the results showed the importance of gender and age for UTI prediction.

Conclusion This study highlighted the potential of machine learning strategies for UTI prediction.

Highlights

- The most informative variables for UTI prediction were suggested.
- An ensemble machine learning approach obtained UTI prediction accuracy of 85.64%.
- Gender and age were important variables for UTI prediction.

Keywords Urinary tract infection, Prediction, Machine learning, Feature extraction

*Correspondence:

Sajjad Farashi
sajjad_farashi@yahoo.com
Hossein Emad Momtaz
hemmtz@yahoo.com

¹Neurophysiology Research Center, Institute of Neuroscience and Mental Health, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran

²Department of Pediatrics, School of Medicine, Ekbatan Hospital, Hamadan University of Medical Sciences, Hamadan, Iran

³Urology and Nephrology Research Center, Avicenna Institute of Clinical Sciences, Avicenna Health Research Institute, Hamadan University of Medical Sciences, Hamadan, Iran



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

Urinary tract infection (UTI) is a common health-threatening issue that affects millions of people each year [1]. Furthermore, the majority of microbiological screenings are associated with cases with possible UTIs [2]. UTI is diagnosed based on mainly urine culture and clinical symptoms. Urine culture is very sensitive to the collection method and final report needs 24–72 h to be accessible. In addition, clinical symptoms alone are not reliable for an accurate prediction [3]. In this regard, alternative fast methods are required for UTI prediction. Currently, the lack of fast, accurate and reliable tools for UTI prediction causes the overuse of antibiotics which contributes to the development of bacterial resistance to antibiotic agents.

One possible solution for predicting UTI is developing artificial intelligence (AI) systems where artificial machines are trained using extracted information from available data. AI systems are effective solutions for reducing the cost and errors caused by human incorrect decisions and are fast due to the capability of computer systems compared with humans [3]. However, the performance of an AI system for predicting UTI depends heavily on identifying the most informative attributes of UTI. In other words, it is important to use the variables with the most discrimination capability between UTI and non-UTI cases.

So far several studies have focused on developing intelligent systems for UTI prediction. In a binary classification study, Taylor et al. released a large dataset in both sample size (more than 80,000 samples) and feature size (more than 200 features) suitable for UTI prediction. They obtained a UTI prediction accuracy of 87.5% with an XGBoost classifier according to full feature set and an accuracy of 85.9% with the same classifier and a reduced feature set. The reduced feature set was selected according to the literature search [4]. Burton et al., using a large dataset (212,554 samples) and selected features by recursive feature elimination strategy obtained a prediction accuracy of 85% using a neural network classifier [2]. The necessity for using different prediction strategies for children (age < 11) and pregnant women was the important result of that study. Heckerling et al. using a small dataset (212 samples) and a combination of artificial neural network and genetic algorithm and a reduced feature set containing urinary frequency, foul urine odor, leukocytes, bacteria, and epithelial cells on urinalysis proposed a model for UTI prediction with an accuracy of 76.4%. In Choi et al. [5], the best discriminative power for UTI prediction was obtained by an XGBoost classifier (85.7%). Gadalla et al. proposed a machine learning tool for predicting UTI in women samples using clinical and immunological factors. According to a recursive feature elimination strategy and SVM and RF classifiers, MMP9,

NGAL, IL-8/CXCL8, and IL-1 β in urine were proposed as the most informative predictors [6]. In Pak et al., a predictive model for UTI in stroke patients was proposed. This study found that age, gender, smoking, alcohol, post-stroke hyperglycemia, the indwelling of urinary catheter, and the coma scale were the most predictive features for UTI prediction. The area under curve for the prediction was 0.8 [7]. In Dedeene et al., a tool based on AI for urine culture test was proposed [8]. The study showed that the multilayer perceptron classifier achieved area under curve of 0.88, while the predictive parameters were urine dipstick, urine cytometry, and demographic features such as age and gender. Furthermore, using a simpler logistic regression model and flow cytometry-based features obtained a lower but acceptable area under curve of 0.86 [8].

According to the literature, different studies suggested various minimal feature sets for accurate UTI prediction. However, the prediction accuracy of proposed models for UTI prediction still needs improvement. The main purpose of the current study was to search for the most informative factors according to an available big dataset for UTI prediction. Another aim of this study was to propose an artificial intelligence system for providing a probability of UTI for blind cases.

Methods

Models were developed using machine learning algorithms in Python programming language (Version 3.8.1), while several packages including numpy (1.26.4), sklearn (1.4.2), Pandas (2.1.4), Scipy (1.11.4), Xgboost(2.0.3), lightgbm(4.3), and deeptables (0.2.6) were incorporated. All other analyses such as statistical analyses, and figure generation were also performed in Python.

Dataset description

In this study, the available dataset prepared by Taylor et al. from a single-center, multi-site study was used [4]. This dataset contains more than 200 variables for each susceptible case who visited four adult emergency departments from March 2013 until May 2016. For each case, information regarding vital signs (body temperature, blood pressure, O₂ saturation, and dependency), laboratory results, and urinalysis results, history of medication, the complaints (including abdominal pain, back pain, altered mental status, dizziness, fatigue, genital problems, fever, flank pain, hematuria, and weakness), demographics, and physical exam findings were collected. This tabular dataset contained both categorical and numerical data. The vital sign variables transformed to a clustered space in which each value was assigned to one cluster according to its level (1: critically low, 2: low, 3: normal, 4: high, 5: critically high). The labeling strategy (UTI vs. non-UTI) was described in Taylor et al. [4]. This dataset contains

more than 80,000 samples. More information about this dataset can be found in [4].

Feature selection

The used dataset for the current study includes more than 200 features. Since the major purpose of the current study was to propose a clinically applicable AI system, such a high-dimensional feature space must be reduced to a lower-dimensional space. This reduces the computational and time complexity of training step of AI system and also makes it feasible for clinical staff to feed the most informative data to the AI system for the prediction of a possible UTI.

To reduce the dimensionality of dataset, feature selection was performed in three successive steps. At first, in manual feature elimination, some features such as race, ethnicity, or language, which are specific to a particular area, were excluded. Some features with a vague definition such as weakness, or psychiatric confusion were also excluded. Furthermore, features which contain many missing values ('not_reported' values in the dataset) were excluded. A 70% threshold for missing values relative to the total sample size was used for ignoring such features. In addition, since this study tries to propose an AI strategy for predicting UTI at admission, features such as O2 saturation, blood pressure and body temperature were considered only at the time of admission and such variables during hospitalization were ignored. For the second round of feature space reduction, five different types of feature selection strategies were used, including filter methods (i.e. information gain, chi-squared, Fisher's score, and mean absolute difference), and also an embedded method (Lasso regularization feature selection). The information gain method calculates the mutual information (MI) or dependency of each feature and the label vector (UTI and non-UTI labels) according to the entropy estimation [9]. A higher value of MI indicates higher dependency (i.e. feature is a representative of the label vector). The chi-squared method computes statistical dependence between each feature and class labels. Fisher score is calculated according to the ratio of between-class and within-class variance and its higher value indicates the more discrimination power of a feature. The mean absolute difference calculates statistical dispersion and returns the average absolute difference between a selected feature and associated class labels [10]. More details about these feature selection methods can be found in supporting materials (see supporting materials, Details of feature selection tools section).

Each method identifies the most informative features from a unique perspective. For practical feasibility, each feature selection method was forced to find the 20 most informative features. After accumulating selected features by different methods, features that were labeled

as informative by at least two feature selection methods were chosen. Finally, for reducing the feature space to a smaller size, the third round of feature selection using wrapper methods was applied. The wrapper feature selection methods including forward, backward, and also recursive feature elimination methods were used to select the most informative subset of features. The final feature set with the acceptable prediction performance was searched according to the classification results. It is worth noting that wrapper methods are very time-consuming when dealing with a high-dimensional data. In this regard, these methods were used after reducing the dimension of data using more time-efficient methods like filter feature selection strategies. To calculate the importance of selected features, the neighborhood component analysis (NCA) was performed [11]. This algorithm is a machine learning strategy which is based on nearest neighbor-based feature weighting according to the maximizing classification accuracy. NCA feature reduction strategy obtained enhanced classification results compared with other dimension reduction algorithms like principal component analysis [12].

Classification

In this study, different types of classifiers were used. Classic classifiers like logistic regression, Naïve Gaussian Bayes classifier, decision tree and random forest classifier, eXtreme Gradient Boosting (XGboost), and light gradient-boosting machines (LightGBM) [13] and also several deep learning-based machines including deep factorization machines (DeepFM) [14], eXtreme deep factorization machine (xdeepFM) [15], wide and deep network (WDN) [16], deep and cross network (DCN) [17], product-based neural networks (PNN) [18], automatic feature interaction learning via self-attentive neural networks (AutoInt) [19], attentional factorization machines (AFM) [20], and feature generation by convolutional neural network (FGCNN) [21] were also used. Data was split randomly into two chunks including train and test where the test size was 25% of the total sample size. A parameter optimization step, using grid search or randomized search algorithms for tuning each classifier was performed before the training step. For parameter optimization, 20% of training data was chosen randomly to avoid overfitting. For training each classifier, a 10-fold cross-validation was used and the holdout samples were used for model validation.

It should be noted that the dataset used in this study is highly imbalanced (80389 samples, in which 62104 samples were labeled as non-UTI). There are many suggestions for preventing biases due to data imbalance. In the current study, some of these suggestions were considered. First, a range of measures such as accuracy, precision, recall, and area under the curve were used for

performance evaluation. Furthermore, algorithms such as random forest and boosting-based techniques such as XGboost and lightGBM that have been used in this study are robust against imbalanced data. In such methods, combining several models that are trained using different data subsets reduces the effects of data imbalance [22]. Another strategy for handling imbalance classes was repeated k-fold cross-validation during the testing procedure. This strategy introduces both randomization and multiple evaluations and in this way reduces the effect of data imbalance. Finally, when the dataset was split into test and train samples, the stratified random sampling was used in which each set contained an equal percentage of samples from different classes.

Test data was used to evaluate the performance of the optimized classifier according to the accuracy, precision, and recall metrics. To check the generalizability of the classifier and prevent issues like imbalanced classes and overfitting, repeated K-fold cross validation ($K=5$) was used. In this strategy the training and testing steps were repeated by including different test and train samples and the mean and standard deviation of results for these repeats were reported. Classifiers were compared according to the area under ROC curve (AUC) in which the AUC value was scaled in 0-100 range (instead of 0–1) to show the small differences between classifiers' performance.

Results

The first feature reduction step (i.e. manual feature elimination step) reduced the size of feature space from 211 to 150. In table S1 (see supporting material), selected features by the second phase of feature selection (i.e. using different filter methods and Lasso regularization method as an embedded feature selection strategy) were shown. By applying filter and embedded feature selection strategies and accumulating the selected features, thirty-two features (feature set No.1) were selected (Table 1). The criterion for selecting feature set No. 1 was that the feature should be proposed by at least two feature selection methods in phase two.

The results of classification using feature set No.1 (32 selected features, as shown in Table 1) and different machine learning strategies were reported in Table 2. Using a K-fold cross-validation ($K=10$), one fold of training samples was used for validation in a repeated manner. The testing samples were used for testing phase.

According to Table 2, the performance of XGboost classifier in terms of classification accuracy, precision, specification, recall and AUC was better than other classifiers. LightGBM, DeepFM and Decision tree classifiers were other top classifier options for further analysis. For the final step of feature selection, three different wrapper methods were used. After third feature selection

strategy, i.e. wrapper feature selection methods, the initial 211-dimensional feature space was reduced to an 18-dimensional space. The selected features by each method were shown in Table S2 (see supporting materials). The final feature set was selected by accumulating the features proposed by wrapper methods. In this regard, eighteen features were proposed as the most discriminative features (feature set No.2) as ua_nitrite, ua_wbc, ua_bili, ua_leuk, ua_urobili, abx, abxUTI, age, chief_complaint, gender, Glucose, Lymphocytes, MPV, Potassium, RDW, ua_blood, ua_clarity, and ua_color. The classification performance for feature set No.2 and qualified classifiers was reported in Table 3.

According to the results in Table 3, when a combination of three classical classifiers including decision tree, XGBoost, and lightGBM with equal weights as an ensemble classifier was used, the maximum value for classification performance was obtained.

Some notes about selected features

In Fig. 1, the percentage of cases with positive nitrite and urobilinogen report, the percentage of cases with large WBC, leucocytes, blood and bilirubin in urine sample, and critically large blood glucose, lymphocytes, potassium, MPV, and RDW reports for different classes (UTI infected vs. non-UTI) were shown. According to Fig. 1, among the most informative features with categorical values (positive vs. negative, or large vs. other levels), urine nitrite, urine WBC, urine leucocyte, urine blood, urine bilirubin, blood glucose, and blood potassium were features in which the percentage in the infected group was larger than the non-infected group. Furthermore, according to Fig. 1, higher percentage of non-UTI cases had critically large blood lymphocytes, RDW, and MPV and urine urobilinogen. This implies that such features possibly reduced following UTI.

Correlation between selected features

To check for any correlation among selected features, correlation matrix is shown in Fig. 2. The categorical variables were coded to a binary space as follows: for ua_nitrite: positive (1) vs. negative (0); for ua_urobili since according to Fig. 1 the number of positive cases was smaller in UTI as compared with non-UTI group: positive (0) vs. negative (1); for ua_wbc, ua_leuk, ua_blood, ua_bili: large (1) vs. other levels (0); for glucose, and potassium: critically large (1) vs. other levels (0); for lymphocyte, MPV, and RDW: critically large (0) vs. other levels (1), since according to Fig. 1 the number of critically large cases was smaller in UTI as compared with the non-UTI group; for ABX and ABX_UTI: yes (1) vs. no (0), and for ua_clarity: not_clear (1) vs. clear (0). Features such as chief complaint and ua_color were not considered for this analysis, since the exact relationship between them

Table 1 The most informative feature set (feature set No.1) for UTI prediction according to the filter and embedded feature selection strategies

Feature name	Description	Number of selection by five feature selection methods
ua_wbc	Number of white blood cells in the urine (large, moderate, small, negative)	5
Blood basophils	Amount of a types of white blood cells called granulocytes	4
chief_complaint	Complaint such as fever, pain, dizziness, fatigue, dysuria, genitourinary problems, weakness, and Hematuria	4
ua_clarity	Clarity of urine (clear or not clear)	4
ua_nitrite	The presence of nitrite in urine (negative or positive)	4
abx	Antibiotic consumption (yes or no)	3
Blood ANC	Absolute Neutrophil Count in blood	3
Blood eosinophils	White blood cells for boosting inflammation	3
fever	Body temperature higher than 37° C (yes or no)	3
Blood lymphocytes	A type of immune cell in the blood	3
Blood monocytes	A type of white blood cell in the blood	3
Blood MPV	Mean platelet volume of blood	3
ua_rbc	Number of red blood cells in the urine (large, moderate, small)	3
abd_pain	Abdominal pain (yes or no)	2
abd_soft	abdomen softening (yes or no)	2
abxUTI	Antibiotic consumption for UTI (yes or no)	2
age	Age of participant	2
arrival	The way that subject was delivered to the emergency unit (Car, EMS, walk-in, or Wheelchair)	2
gender	Male or female	2
Glucose	Blood glucose level	2
Blood MCH	Mean corpuscular hemoglobin	2
Blood potassium	Blood potassium level	2
Blood RDW	Red blood cell distribution width	2
ua_bili	Bilirubin in urine (large, moderate, negative, small)	2
ua_blood	The presence of blood in urine (large, moderate, negative, small)	2
ua_color	The color of urine (amber, colorless,, red, yellow, other)	2
ua_glucose	Urine glucose level (large, moderate, small, negative)	2
ua_ketones	Urine ketones level (large, moderate, small, negative)	2
ua_leuk	Leukocytes in Urine (large, moderate, small, negative)	2
ua_protein	Protein in urine (large, moderate, small, negative)	2
ua_spec_grav	Urine specific gravity (numeric value)	2
ua_urobili	Urobilinogen in the urine (positive or negative)	2

and UTI was unknown. In the used dataset, for both UTI positive and negative cases, the yellow color urine was dominant (85.89% and 88.77% of UTI positive and negative cases, respectively, while the frequency of other urine colors was negligible). In addition, the most frequent chief complaint in both UTI and non-UTI group was abdominal pain (23.73% and 31.58%, respectively) and other symptoms were not dominant. Furthermore, there was no standard threshold for the age variable regarding UTI incidence. Therefore, it was impossible to map these features into a binary representation. For this analysis, only cases were considered in which all 14 above-mentioned variables were reported (48761 samples). Pearson's correlation was used for calculating the correlation coefficient (R) and associated p-value.

In Table 4 the weight (importance) of selected features according to the neighborhood component analysis was reported.

Sensitivity of classifier performance to age and gender

The feature selection stage showed that age and gender were important factors for UTI prediction. To check how the performance of the prediction system was affected by age or gender, the prediction capability for different age spans (18–40, 40–60, and >60 years of old) and different genders (male or female) was investigated. In Figs. 3 and 4, ROC curve for UTI prediction for different age spans and different genders was shown. For these analyses, an ensemble learning technique by combining decision tree, XGboost, and lightGBM estimators with equal weight was used as was shown in Table 3 this obtained the best

Table 2 Result of classification with a K-fold cross-validation (K = 10) for training step and repeated K-fold cross-validation (K = 5) for testing phase. Reported values are mean (standard deviation). The feature set was as table 1

Classifier	Validation phase				Testing phase				
	Accuracy (%)	Precision (%)	Recall (%)	AUC	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	AUC
Classic classifiers									
Logistic regression	83.09 (0.29)	70.36 (1.14)	44.26 (1.28)	83.70 (0.56)	83.24 (0.23)	94.67 (0.17)	71.18 (0.78)	44.54 (0.96)	83.83 (0.25)
Naïve Gaussian Bayes	80.78 (0.40)	59.24 (1.10)	49.22 (1.15)	81.40 (0.73)	80.75 (0.10)	90.03 (0.08)	59.54 (0.30)	49.46 (0.12)	81.42 (0.10)
Decision tree	84.99 (0.34)	75.98 (1.79)	49.76 (2.16)	86.98 (0.52)	84.91 (0.28)	95.38 (0.52)	76.10 (1.42)	49.48 (1.50)	86.96 (0.23)
Random forest	84.08 (0.36)	75.30 (2.65)	45.18 (3.89)	85.18 (0.51)	84.16 (0.19)	95.97 (0.11)	76.49 (0.55)	44.28 (0.27)	85.16 (0.17)
XGboost	85.92 (0.37)	75.63 (1.20)	56.28 (1.20)	88.92 (0.49)	86.05 (0.23)	94.76 (0.18)	75.97 (0.55)	56.37 (0.67)	88.87 (0.14)
LightGBM	85.91 (0.37)	75.24 (1.07)	56.62 (1.32)	88.78 (0.53)	85.78 (0.24)	94.48 (0.20)	75.09 (0.66)	56.32 (0.32)	88.81 (0.14)
Deep learning networks									
DeepFM	83.43 (1.19)	78.83 (6.52)	39.40 (12.24)	86.75 (0.69)	85.18 (0.47)	94.49 (1.32)	74.26 (2.95)	53.33 (4.27)	87.69 (0.55)
WDN	83.02 (0.67)	76.85 (3.61)	37.05 (6.86)	85.08 (0.52)	83.40 (0.39)	96.30 (1.11)	76.87 (2.78)	40.18 (4.98)	85.33 (0.16)
xdeepFM	83.18 (0.58)	78.47 (2.88)	36.02 (5.33)	85.16 (0.32)	83.13 (0.73)	96.17 (1.29)	76.32 (4.32)	39.31 (7.43)	85.44 (0.21)
DCN	79.52 (1.15)	89.97 (3.54)	11.19 (6.70)	85.05 (0.39)	82.44 (0.61)	82.44 (0.61)	82.55 (2.84)	29.24 (4.82)	85.29 (0.33)
PNN	83.34 (0.33)	78.07 (2.77)	37.68 (3.76)	83.73 (0.15)	83.44 (0.45)	96.19 (0.56)	76.13 (1.70)	41.22 (2.39)	85.38 (0.26)
AutoInt	60.52 (25.61)	34.65 (31.74)	43.82 (44.58)	77.99 (8.17)	67.19 (22.29)	74.76 (38.21)	42.60 (27.61)	41.68 (38.38)	77.77 (7.99)
AFM	77.36 (0.31)	12.00 (32.50)	0.01 (0.04)	80.76 (2.11)	78.88 (1.91)	99.11 (1.15)	31.42 (38.51)	10.57 (13.09)	82.39 (1.42)
FGCNN	83.68 (0.36)	72.56 (1.79)	45.79 (2.99)	85.49 (0.38)	83.99 (0.13)	94.47 (0.75)	72.12 (1.68)	48.34 (2.87)	85.66 (0.24)

Table 3 Classification results for the selected classifiers according to feature set No.2

Classifier	Validation phase (10-fold cross validation)				Testing phase (repeated 5-fold cross validation)				
	Accuracy (%)	Precision (%)	Recall (%)	AUC	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	AUC
Decision tree	84.98 (0.45)	75.52 (2.03)	50.49 (1.94)	86.97 (0.64)	85.03 (0.22)	95.57 (0.49)	76.45 (1.48)	49.04 (1.99)	86.88 (0.21)
XGboost	85.66 (0.34)	75.30 (1.07)	55.01 (1.20)	88.35 (0.46)	85.58 (0.13)	94.70 (0.19)	75.24 (0.40)	54.63 (0.96)	88.37 (0.11)
LightGBM	85.76 (0.38)	74.95 (1.25)	55.91 (1.11)	88.55 (0.50)	85.80 (0.15)	94.66 (0.22)	75.73 (0.94)	56.04 (0.54)	88.56 (0.19)
DeepFM	84.20 (0.61)	77.64 (4.33)	43.71 (7.11)	87.17 (0.49)	85.00 (0.07)	95.40 (0.73)	76.01 (1.70)	49.43 (2.99)	87.49 (0.15)
Combined classifier (equal weight)	85.70 (0.37)	76.01 (1.27)	54.22 (1.53)	88.52 (0.50)	85.64 (0.20)	94.89 (0.20)	75.86 (0.75)	54.37 (0.24)	89.53 (0.25)

discrimination. The combination used a voting scheme to make the final prediction.

Comparison with other methods

One of the main challenges in UTI prediction is how to determine the minimal feature set required for a fast and accurate prediction. In this regard, different studies tried to find the most informative features according to their dataset. In Table 5, the prediction capability of several studies according to different feature sets was compared with the results obtained by this study. Since various strategies can be applied for implementing machine learning algorithms (such as different data splitting strategies, different cross-validation folds, various classifier implementation, and optimization methods, and etc.), the prediction capability of these feature sets was estimated by the same classification strategy that was used

in our study and the same dataset that was released by Taylor et al. [4].

In Taylor et al. [4], a minimal feature set including age, gender, urine analysis leukocytes, urine nitrite, urine WBC, urine bacteria, urine blood, urine epithelial cells, history of UTI, and dysuria was proposed as the most informative features. These features were selected according to a literature review. In Choi et al. [5], the top ten variables for UTI prediction were reported to be urine bacterial count, monocyte count, WBC count, lymphocyte count, urinary WBC count, specific gravity, diastolic blood pressure, systolic blood pressure, age, mean blood pressure, and C-reactive protein. All these features were included in the dataset of Taylor et al. [4]. The result of UTI prediction based on the above-mentioned features, our developed machine learning strategy and Taylor et al. dataset [4] was reported in Table 5.

Comparison between most informative features of UTI and non-UTI cases

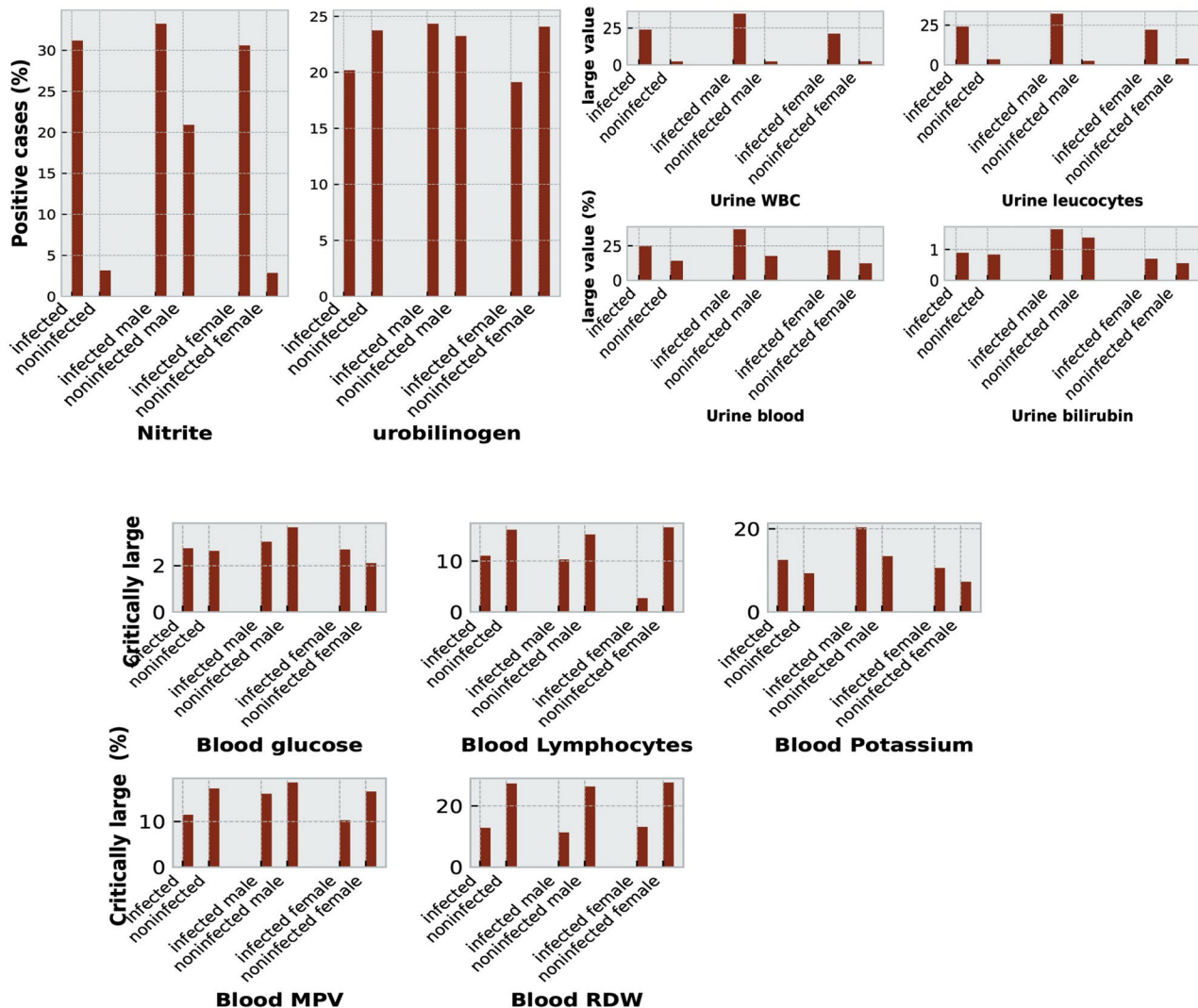


Fig. 1 Comparison between UTI-infected and non-UTI (non-infected) cases, considering the most informative features

In Burton et al. [2], which used an independent dataset compared to the current study and Taylor et al. [4], 21 features including WBC count, bacterial count, age, epithelial cell count, RBC count, number of positive cultures to date, pyuria (no RBC), pregnancy, inpatient, gender, persistent/recurrent, infection, number of positive cultures month prior, positive for nitrates, renal inpatient/outpatient, pre-operative patient, acute kidney disease, immunocompromised, number of positive cultures week prior, multiple sclerosis, offensive smell, haematuria (no WBCs) were selected as the most informative for UTI prediction. The best performance for UTI prediction for this feature set was obtained by the XGboost classifier (AUC=0.91). However, since the dataset of Burton et al. was not accessible, it was not possible to compare the

potential of the proposed feature set for UTI prediction with other methods.

Discussion

The current study suggested urine WBC, urine nitrite, urine leucocyte, and urine clarity as the most informative biomarkers for UTI prediction (see Table 4). Positive nitrite in urine due to the contribution of bacteria to change nitrates to nitrites in the urine might be considered a possible sign of infection in the urinary tract. However, a previous study showed that analysis of solely urine nitrite, WBCs, and leucocytes as a fast strategy for UTI-suspected cases [23] was subject to low sensitivity and specificity [24]. This might be due to the fact that even though higher levels of WBCs and positive nitrite are potential signs of UTI, they are not specific

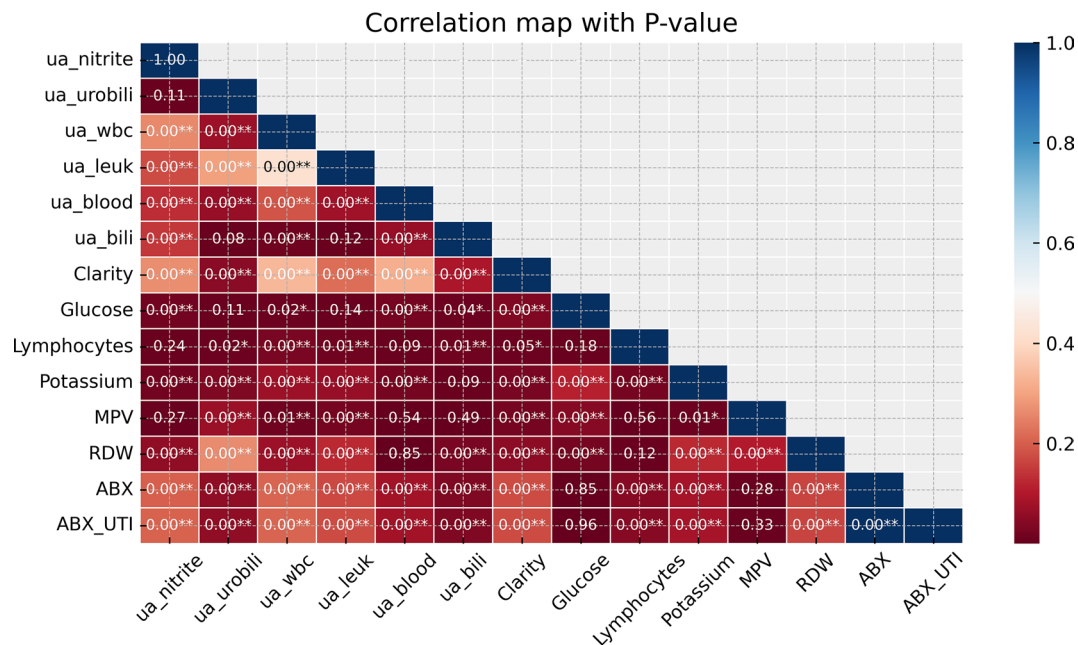


Fig. 2 Correlation matrix of selected variables with categorical values. The color indicates the correlation value and each number indicates the p-value (*: $p < 0.05$, **: $p < 0.01$)

Table 4 Weight of features according to the neighborhood component analysis

Features	weight	Importance score (%)
ua_wbc	0.961	22.41
ua_nitrite	0.926	21.61
ua_leuk	0.842	19.63
ua_clarity	0.545	12.72
MPV	0.156	3.63
Lymphocytes	0.151	3.53
ua_blood	0.146	3.40
ua_color	0.131	3.05
ua_bili	0.098	2.28
ua_urobili	0.076	1.78
abx	0.061	1.42
Age	0.057	1.34
Gender	0.047	1.09
abxUTI	0.048	1.04
Glucose	0.028	0.66
RDW	0.012	0.28
Potassium	0.005	0.13
chief_complaint	0	0

characteristics of UTI. For example, other conditions such as kidney stones, fever, interstitial nephritis, or pelvic problems may trigger the increase of WBC count, or in patients with gastroenteritis, the urinary nitrite significantly increases [25]. The lack of specificity of these factors for UTI was evident in results depicted in Fig. 1 in which a portion of non-infected (non-UTI) samples showed positive nitrite or large levels of WBCs and leucocytes.

Another useful biomarker for bacterial inflammation detection is the neutrophil to lymphocyte ratio, however, it is not specific for UTI and several diseases such as pneumonia, cancers or heart diseases affect such a biomarker [26]. Positive urine bilirubin which was supposed to be a potent biomarker (Tables 1 and 4), is highly correlated with unusual liver functions [27], therefore such an indicator is not the specific biomarker for UTI. The small weight for such a feature (Table 4) and the relatively large non-UTI samples with large bilirubin levels (Fig. 1) showed that bilirubin could not be considered as a potential biomarker for UTI.

Literature suggested a positive correlation between diabetes (higher level of blood glucose) and UTI. Diabetes may damage the nerves and weaken the immunity system and in this way increases the risk of bacterial infections. High blood levels may also provide nutrients for bacteria [28]. This might justify why blood glucose was proposed in the current study as a biomarker for UTI prediction. In line with the results of the current study, MPV as an indicator of the function of platelets, was suggested by another study as a potential biomarker for UTI [29]. MPV during mild inflammation shows an increased rate, while for severe inflammations it decreases due to the consumption of platelets [30]. The direction of MPV change is a controversial issue between studies. Some studies introduced MPV as a positive acute phase reactant and reported its increased level [31], while some other studies considered MPV as a negative acute phase reactant and reported its decrement after inflammation [30]. According to the dataset used in this study, the percentage of

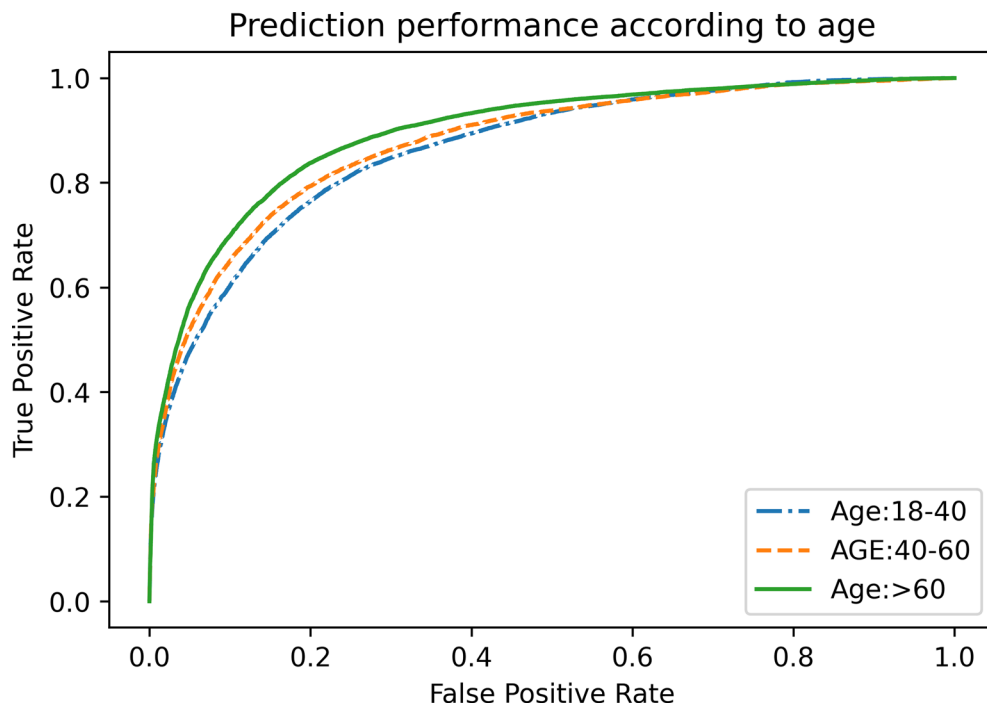


Fig. 3 ROC curve for comparison of predictor performance for different age spans. AUC was 86.42 (0.17) (Age: 18–40), 87.71 (0.22) (Age: 40–60), and 89.42 (0.25) (Age > 60), respectively

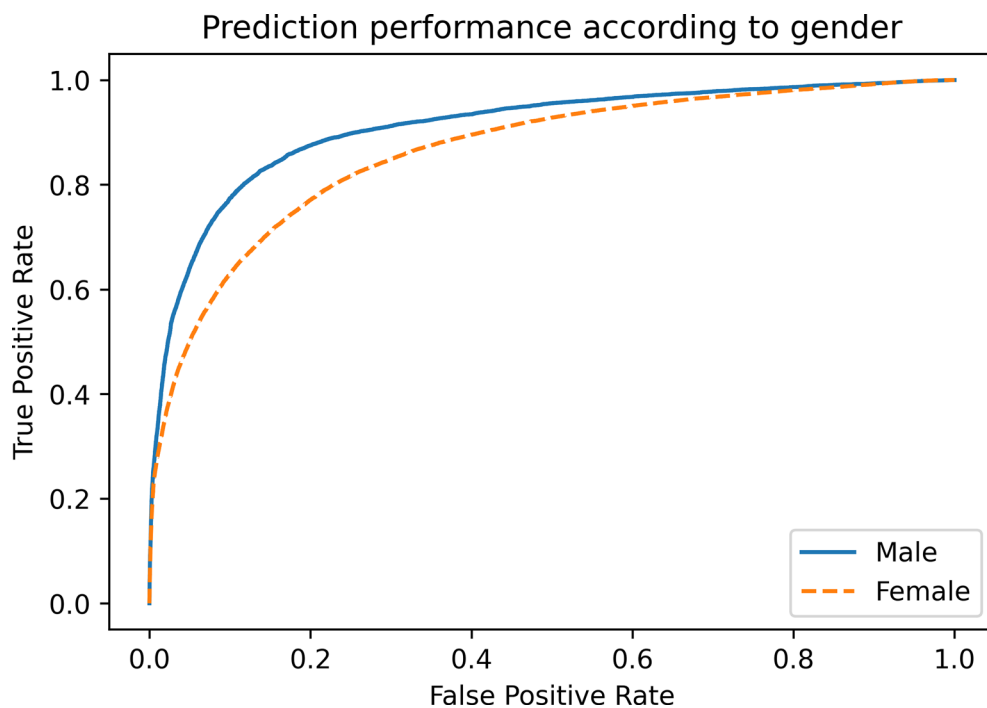


Fig. 4 ROC curve for comparison of predictor performance according to the gender. AUC was 91.42 (0.32) and 86.16 (0.23) for males and females, respectively

UTI-labeled subjects with large levels of MPV was lower compared with non-UTI samples (Fig. 1).

The urine clarity which was proposed as a potential biomarker in the current study (Table 4), is another

indicator that usually is used for UTI prediction. However, previous results implied that visual inspection of urine cannot be a sufficient indicator for UTI prediction [32]. Furthermore, previous studies suggested higher

Table 5 Comparison between UTI prediction capability of different feature sets. The same classification strategy and the same dataset were used for all feature sets

	Accuracy	Specificity	Precision	Recall	AUC*
This work (XGboost)	85.58 (0.13)	94.70 (0.19)	75.24 (0.40)	54.63 (0.96)	88.37 (0.11)
This work (Com-bined classifier)	85.64 (0.20)	94.89 (0.20)	75.86 (0.75)	54.37 (0.24)	88.53 (0.25)
Taylor et al. (XGboost)	84.39 (0.20)	94.29 (0.21)	72.41 (0.81)	50.84 (0.84)	86.26 (0.22)
Taylor et al. (combined classifier)	84.44 (0.23)	95.26 (0.28)	74.85 (0.50)	47.77 (1.47)	86.34 (0.56)
Choi et al. (XGboost)	83.95 (0.24)	94.05 (0.13)	71.14 (0.36)	49.70 (0.66)	85.43 (0.30)
Choi et al. (Com-bined classifier)	84.27 (0.26)	94.44 (0.18)	72.43 (0.48)	49.66 (0.63)	85.82 (0.36)

*To show minor differences AUC was scaled up to 0-100 range

sensitivity and specificity of WBC count than nitrite when compared with urine culture [33]. This is in accordance with the higher importance score of urine WBC as compared with nitrite, however, other factors like blood glucose level or pregnancy might influence such an outcome [33]. The results of this study proposed RDW as one of the potential biomarkers of UTI prediction. Previous studies reported the correlation between RDW and systemic inflammation [34]. The results reported in Fig. 1 indicated a lower percentage of UTI cases had large levels of RDW as compared with non-UTI samples. Another potent factor that was highlighted in the current study for UTI prediction was potassium level (Table 4). Several studies reported the ionic abnormalities among inpatients with UTI [35, 36]. Ionic abnormality may be due to poor feeding, increased sweating, or vomiting in UTI conditions [37]. Low potassium level (hypokalemia) was commonly observed in UTI cases [38], while for pyelonephritis cases (a type of urinary tract infection with infected kidneys), the increased level of circulating potassium was observed compared with the control group [39]. The increased potassium level is in accordance with the analysis of the used dataset (see Fig. 1) in which the prevalence of critically large blood potassium was observed for the UTI group.

According to Fig. 2, there was no strong correlation ($R > 0.5$) between features except for ABX (antibiotic usage) and ABX_UTI (antibiotic usage for UTI), which are highly correlated ($R = 1$). This may be due to the fact that people who use antibiotics to a large extent are more susceptible to UTI due to antibiotic resistance. It should be noted that the prediction performance of the classifier by removing one of ABX or ABX_UTI variables degraded

slightly (not reported in the manuscript). Therefore, it is suggested to use both of them. Furthermore, a moderate correlation ($R = 0.42$, $p < 0.01$) was observed between leukocyte esterase (ua_leuk) and white blood cell counts in urine (ua_wbc). This correlation is reasonable since leukocyte esterase is an enzyme found in white blood cells. The non-strong correlations indicate that selected features do not share common information regarding UTI.

The overall classification accuracy of this work was 85.64% (the result for larger AUC) which was obtained by an ensemble classifier using the combination of decision tree, XGboost and lightGBM classifiers with equal weights (Table 3). However, this performance degraded when UTI prediction was done for female samples or younger cases (Figs. 3 and 4). These results showed the importance of gender and age for developing machine learning strategies for UTI prediction. In addition, when the proposed feature set of this work was compared with other studies (Table 5), it was clear that the combination of classifiers and the selected feature set obtained better accuracy and an enhanced trade-off between type 1 and type 2 errors for UTI prediction.

Study limitations

There are several limitations to the current study. Since the dataset released by Taylor et al. was used in the current study, many limitations of the Taylor et al. study regarding the nature of data such as “no universally accepted threshold for a positive urine culture”, “the absence of a ground truth data or external validation data”, “the necessity for urine culture result”, and finally “exclusion of unstructured data such as clinical notes” are still the limitation of the current study. Especially, the absence of an external validation dataset limited us to check the generalizability of the obtained results. Furthermore, for some of the reported most informative features, a clustering approach was used to allocate the variable into one distinct level (i.e. 0: unknown, 1: critically low, 2: low, 3: normal, 4: high, 5: critically high). Since the details of the clustering approach were unknown to us, it was not possible to talk about the threshold level of discriminative variables.

Conclusion

UTI is a frequent problem in different societies. The reliable prediction of UTI in a quick time prevents unnecessary antibiotic use for non-UTI cases. Machine learning strategies according to the related UTI data are interesting tools for developing prediction systems for UTI prediction. The current study used an available dataset which contained several features including urinalysis, blood tests, and vital signs, demographic and observational data for UTI suspected cases. Finding the most informative features as well as developing an intelligent

system for the predicting UTI were the main purposes of this study. This study showed the potential of machine learning strategies for prediction of UTI according to laboratory and urinalysis results. Furthermore, eighteen features with maximum discrimination capability were proposed. In addition, the results showed that age and gender were two factors that affected UTI prediction.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-024-02819-2>.

Supplementary Material 1

Acknowledgements

Authors would like to thank vice-chancellor for research and technology for all support for the current work.

Author contributions

Sajjad Farashi and Hossein Emad Momtaz performed systematic searches. Sajjad Farashi performed analyses and wrote the manuscript. Sajjad Farashi, and Hossein Emad Momtaz discussed the obtained results and finalized the draft.

Funding

This work was supported by Hamadan University of Medical Sciences, Deputy of research and technology (No. 140308157007/ IR.UMSHA.REC. 1403.569).

Data availability

The data used in this study is fully accessible from <https://doi.org/10.1371/journal.pone.0194085>.

Declarations

Ethical approval

Not applicable.

Consent for publication

Not applicable.

Consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 September 2024 / Accepted: 13 December 2024

Published online: 09 January 2025

References

1. Stamm WE, Norrby SR. Urinary tract infections: disease panorama and challenges. *J Infect Dis*. 2001;183(Suppl 1):S1–4. <https://doi.org/10.1086/318850>.
2. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Med Inf Decis Mak*. 2019;19(1):171. <https://doi.org/10.1186/s12911-019-0878-9>.
3. Goździkiewicz N, Zwolińska D, Polak-Jonkisz D. The Use of Artificial Intelligence algorithms in the diagnosis of urinary tract Infections-A literature review. *J Clin Med*. 2022;11. <https://doi.org/10.3390/jcm11102734>.
4. Taylor RA, Moore CL, Cheung K-H, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. *PLoS ONE*. 2018;13(3):e0194085. <https://doi.org/10.1371/journal.pone.0194085>.
5. Choi MH, Kim D, Park Y, Jeong SH. 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(1):10–7. <https://doi.org/10.1016/j.jiph.2023.10.021>.
6. Gadalla AAH, Friberg IM, Kift-Morgan A, Zhang J, Eberl M, Topley N, et al. Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms. *Sci Rep*. 2019;9(1):19694. <https://doi.org/10.1038/s41598-019-55523-x>.
7. Pak I-H, Han S-R, Sin C-H, Kim H-S, Rim U-R. The development of simple Scoring System to predict urinary tract infection (UTI) in patients with stroke. *Int J Endocrinol*. 2024;2024(1):2512824. <https://doi.org/10.1155/2024/2512824>.
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. Kozachenko LF, Leonenko NN. Sample Estimate of the Entropy of a Random Vector. *Probl Peredachi Inf*. 1987;23(2):9–16.
10. Dodge Y. The concise encyclopedia of statistics. New York: Springer; 2010.
11. Yang W, Wang K, Zuo W. Neighborhood Component feature selection for high-dimensional data. *J Comput*. 2012;7:161–8. <https://doi.org/10.4304/jcp.1161-168>.
12. Kumar Agrawal A, Chakraborty G. Neighborhood component analysis to leverage the class label information during feature selection to enhance the damage classification performance. *Structures*. 2023;57:105174. <https://doi.org/10.1016/j.jistruc.2023.105174>.
13. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. editors. LightGBM: a highly efficient gradient boosting decision Tree. *Neural Information Processing Systems*; 2017.
14. Guo H, Tang R, Ye Y, Li Z, He X. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv Preprint arXiv:170304247*. 2017.
15. Lian J, Zhou X, Zhang F, Chen Z, Xie X, Sun G, editors. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*; 2018. pp. 1754–63.
16. Cheng H-T, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H et al. Wide & Deep Learning for Recommender Systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 2016. pp. 7–10. <https://doi.org/10.1145/2988450.2988454>.
17. Wang R, Fu B, Fu G, Wang M. Deep & Cross Network for Ad Click Predictions. *Proceedings of the ADKDD'17*. 2017.
18. Qu Y, Cai H, Ren K, Zhang W, Yu Y, Wen Y, Wang J, editors. Product-Based Neural Networks for User Response Prediction. *2016 IEEE 16th International Conference on Data Mining (ICDM)*; 2016 12–15 Dec. 2016. pp. 1149–54. <https://doi.org/10.1109/ICDM.2016.0151>.
19. Song W, Shi C, Xiao Z, Duan Z, Xu Y, Zhang M, Tang J, editors. AutoInt: Automatic feature interaction learning via self-attentive neural networks. *Proceedings of the 28th ACM international conference on information and knowledge management*; 2019. pp. 1161–70.
20. Xiao J, Ye H, He X, Zhang H, Wu F, Chua T-S. Attentional factorization machines: learning the weight of feature interactions via attention networks. *arXiv Preprint arXiv:170804617*. 2017.
21. Liu B, Tang R, Chen Y, Yu J, Guo H, Zhang Y, editors. Feature generation by convolutional neural network for click-through rate prediction. *The World Wide Web Conference*; 2019. pp. 1119–29.
22. Khan AA, Chaudhari O, Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *Expert Syst Appl*. 2024;244:122778. <https://doi.org/10.1016/j.eswa.2023.122778>.
23. Appenheimer AB, Ford B. Urine dipstick: urinary nitrites and Leukocyte esterase – dipping into Murky Waters. In: Sharp VJA, Antes LM, Sanders ML, Lockwood GM, editors. *Urine tests: a case-based guide to clinical evaluation and application*. Cham: Springer International Publishing; 2020. pp. 97–115.
24. Williams GJ, Macaskill P, Chan SF, Turner RM, Hodson E, Craig JC. Absolute and relative accuracy of rapid urine tests for urinary tract infection in children: a meta-analysis. *Lancet Infect Dis*. 2010;10(4):240–50. [https://doi.org/10.1016/S1473-3099\(10\)70031-1](https://doi.org/10.1016/S1473-3099(10)70031-1).
25. Wanchu A, Khullar M, Sud A, Deodhar SD, Bamberg P. Elevated urinary nitrite and citrulline levels in patients with rheumatoid arthritis. *Inflammopharmacology*. 1999;7(2):155–61. <https://doi.org/10.1007/BF02918387>.
26. Han SY, Lee IR, Park SJ, Kim JH, Shin JI. Usefulness of neutrophil-lymphocyte ratio in young children with febrile urinary tract infection. *Korean J Pediatr*. 2016;59(3):139–44. <https://doi.org/10.3345/kjp.2016.59.3.139>.

27. Foley KF, Wasserman J. Are unexpected positive dipstick urine bilirubin results clinically significant? A retrospective review. *Lab Med*. 2014;45(1):59–61. <https://doi.org/10.1309/LML85XX1QBERCIRN>.
28. Ahmed AE, Abdelkarim S, Zenida M, Baiti MAH, Alhazmi AAY, Alfaifi BAH, et al. Prevalence and Associated Risk factors of urinary tract infection among Diabetic patients: a cross-sectional study. *Healthc (Basel)*. 2023;11(6). <https://doi.org/10.3390/healthcare11060861>.
29. Akya A, Rostami-Far Z, Chegene Lorestani R, Khazaei S, Elahi A, Rostamian M, et al. Platelet indices as useful indicators of urinary tract infection. *Iran J Ped Hematol Oncol*. 2019;9(3):159–65.
30. Tanju C, Ekrem G, Emel AB, Nur A. Mean platelet volume as a negative marker of inflammation in children with rotavirus gastroenteritis. *Iran J Pediatr*. 2014;24(5):617.
31. Albayrak Y, Albayrak A, Albayrak F, Yildirim R, Aylu B, Uyanik A, et al. Mean platelet volume: a new predictor in confirming acute appendicitis diagnosis. *Clin Appl Thromb Hemost*. 2011;17(4):362–6. <https://doi.org/10.1177/1076029610364520>.
32. Bulloch B, Bauscher JC, Pomerantz WJ, Connors JM, Mahabee-Gittens M, Dowd MD. Can urine clarity exclude the diagnosis of urinary tract infection? *Pediatrics*. 2000;106(5):e60–e. <https://doi.org/10.1542/peds.106.5.e60>.
33. Mohanna AT, Alshamrani KM, SaemAldahar MA, Kidwai AO, Kaneetah AH, Khan MA, Mazraani N. The sensitivity and specificity of White Blood cells and Nitrite in Dipstick Urinalysis in Association with urine culture in detecting infection in adults from October 2016 to October 2019 at King Abdulaziz Medical City. *Cureus*. 2021;13(6):e15436. <https://doi.org/10.7759/cureus.15436>.
34. Ma W, Mao S, Bao M, Wu Y, Guo Y, Liu J, et al. Prognostic significance of red cell distribution width in bladder cancer. *Transl Androl Urol*. 2020;9(2):295–302.
35. Park SJ, Oh YS, Choi MJ, Shin JI, Kim KH. Hyponatremia may reflect severe inflammation in children with febrile urinary tract infection. *Pediatr Nephrol*. 2012;27:2261–7. <https://doi.org/10.1007/s00467-012-2267-9>.
36. Winberg J. Renal function studies in infants and children with acute, nonobstructive urinary tract infections. *Acta Paediatr*. 1959;48:577–89.
37. Bertini A, Milani GP, Simonetti GD, Fossali EF, Faré PB, Bianchetti MG, Lava SAG, Na+, K+, Cl–, acid–base or H2O homeostasis in children with urinary tract infections: a narrative review. *Pediatr Nephrol*. 2016;31(9):1403–9. <https://doi.org/10.1007/s00467-015-3273-5>.
38. Shen A-L, Lin H-L, Lin H-C, Tseng Y-F, Hsu C-Y, Chou C-Y. Urinary tract infection is associated with hypokalemia: a case control study. *BMC Urol*. 2020;20(1):108. <https://doi.org/10.1186/s12894-020-00678-3>.
39. Watanabe T. Hyponatremia and hyperkalemia in infants with acute pyelonephritis. *Pediatr Nephrol*. 2004;19:361–2. <https://doi.org/10.1007/s00467-003-1393-9>.

Publisher's note

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