A hybrid feature selection approach for urinary tract infection detection and prediction in IoT-Fog environment

Anisha Nagpal | Munish Sabharwal | Rohit Tripathi

Abstract: Urinary Tract Infections (UTIs) are a prevalent health concern experienced by millions of individuals worldwide and have a significant impact on overall well-being. The clinical presentation of UTIs varies depending on their location and severity. Common symptoms include discomfort during urination, increased frequency, a compelling urge to urinate, discomfort in the lower abdomen, and the presence of blood in the urine (hematuria). Severe cases may be accompanied by fever, pain in the flank region, and other systemic symptoms, indicative of an upper UTI. Precise diagnosis of UTIs is achieved through clinical evaluation and laboratory tests. The primary aim of this research is to utilize appropriate Machine Learning (ML)-based algorithms to predict Urinary Tract Infections (UTIs). The ultimate goal is to develop a predictive model that can be effectively implemented in a smart toilet system. By achieving this objective, this research seeks to provide an innovative and pragmatic solution for UTI prediction, harnessing the potential of ML algorithms in IoT-Fog settings to enhance healthcare and promote public health. This paper introduces a hybrid approach that combines feature selection methods and employs the Guided Regularized Random Forest (GRRF) classification algorithm to aid in the diagnosis of UTIs. A UTI dataset was created using data from routine examinations and definitive diagnostic outcomes for UTI patients. Dimensionality reduction was carried out using Principal Component Analysis (PCA), while feature selection was performed using K-best and Lasso CV techniques. Through the proposed strategy, this study achieved a remarkable accuracy rate of 98.8% and a precision rate of 98.90% in UTI identification. Future UTI prevention and treatment plans must be optimized via further research and ongoing efforts to overcome antibiotic resistance.

Keywords: urinary tract infection, guided regularized random forest, hybrid feature selection, machine learning

1. Introduction

ICT advancements have emerged as crucial assets in diverse sectors such as healthcare, logistics, and agriculture, offering efficient solutions. The Internet of Things (IoT) has played a significant role in propelling these ICT innovations. In the healthcare industry, this has led to better resource management and the widespread availability of healthcare services. In recent times, IoT has emerged as an increasingly attractive and revolutionary technology, playing a crucial role in diverse fields, including healthcare, house automation, smart cities, wearable devices, and more. IoT has a profound impact on the healthcare industry, enabling various applications of smart healthcare (Bansal, 2022; Lin, 2019).

In the forthcoming years, the significance of fog computing is projected to be of paramount importance in addressing the ever-growing need for instantaneous services. As a platform, fog computing offers increased storage capacity, real-time computational power, and network services, effectively bridging the gap between data centers and end-users. The integration of IoT-Fog computing facilitates the execution of numerous time-sensitive data and services, including emergency health services and medical diagnosis. Smart healthcare applications encompass intelligent patient monitoring, wireless health tracking, and mobile healthcare services. Many cities are actively pursuing the concept of smart city healthcare, utilizing conventional equipment and devices that integrate healthcare resources with smart solutions. This convergence of technologies is poised to revolutionize healthcare services, enabling more efficient and responsive healthcare delivery for the benefit of individuals and communities (Alshamrani, 2022).

Machine learning models for predictions are proving to be invaluable tools in clinical practice, as they offer improved guidelines for decision-making in personalized patient care. These models have the capability to self-diagnose a wide range of diseases, aligning their assessments with established clinical guidelines. This integration of machine learning in healthcare empowers medical professionals to make more informed and precise decisions, ultimately leading to better outcomes for individual patients. In addition, IoT has been integrated with ML techniques to monitor the well-being and health of...
individuals affected by dementia. This combined model aids in delivering more effective preventive care, ultimately reducing the need for hospitalization (Enshaeifar, 2019, Gadalla, 2019).

The primary objective of this current research is to formulate an innovative algorithm grounded with Machine Learning (ML) techniques within an IoT-Fog environment, specifically applied through a smart toilet system. Additionally, the study aims to enhance the accuracy rate of UTI prediction, as this factor holds significant importance in ensuring better patient care.

Furthermore, the study recognizes the availability of various IoT devices with advanced capabilities in acquiring specific urine parameters and detecting UTIs. Particularly, IoT encompasses internet-enabled sensors capable of gathering comprehensive data and transmitting it to remote locations. Moreover, these devices employ easily accessible technologies, making them suitable for integration into a smart toilet system. By utilizing such IoT devices within the smart toilet system, the study anticipates a promising avenue for effective UTI detection and monitoring, thereby contributing to improved healthcare outcomes and overall patient well-being (Ijaz, 2021).

Numerous people across the world suffer from urinary tract infections (UTIs), which are a prevalent medical problem (Bankar, 2021). UTIs are brought on by bacteria that enter and attack the urinary tract, which consists of the bladder, urethra, ureters, and kidneys (Verma, 2023). These bacteria are typically from the digestive system i.e Escherichia coli. The signs and symptoms of this infection can range from unpleasant to possibly dangerous. It can affect people of any age or gender identity, but women are more probable than men to experience them due to the female urethra’s being smaller, which makes it simpler for bacteria to get into the urinary system (Wong, 2023). The use of urinary catheters, anomalies of the urinary tract, pregnancy, menopause, a compromised immune system, and certain medical disorders including diabetes are additional risk factors. A medical expert frequently obtains a urine sample to diagnose UTIs in order to check for bacteria or white blood cells (Silva, 2022). An antibiotic course is typically prescribed as part of the treatment to eradicate the disease. To ensure full elimination of the germs and avoid repeated sickness, it is imperative to finish the entire prescribed course of medicines. UTIs can be prevented by taking proactive steps, such as consuming lots of water, urinating regularly, and clearing the bladder both before and after sexual activity (Baijwan, 2023). Further examination by a medical professional may be required for people who have recurring UTIs in order to determine the root cause and create a personalized preventative strategy. The infection can be obstructing but with the right medical care, they are typically manageable (Aggarwal, 2022). People can reduce their risk and seek prompt medical attention when necessary by being aware of the causes, signs, and preventative measures linked to infection (Sulis, 2022).

The early and efficient management of those with urinary tract infections depends critically on the detection and prediction of these conditions. Machine learning and data-driven methodologies have recently demonstrated considerable potential in supporting medical practitioners in identifying and forecasting UTIs (Bijiani, 2022).

A hybrid features selection method is one such strategy that utilizes the effectiveness of various feature selection techniques in order to increase UTI detection and predicting reliability and precision (Gehringer, 2021). In this study, a novel mixed feature choice methodology for UTI identification and forecasting is presented. The suggested approach makes use of the advantages of several algorithms for choosing features to extract the most useful and pertinent features from a wide range of input parameters. The hybrid technique when compared to conventional feature selection methods, the combination has a number of benefits (Hateet 2022). It can get over the drawbacks associated with separate procedures and provide a more thorough evaluation of the feature space by integrating various approaches (Behzadi, 2019).

The generalization of the generated models is improved, and the risk of over fitting is decreased. Additionally, the hybrid feature selection technique is effective in handling duplicate or insignificant characteristics and high-dimensional information, enhancing computational effectiveness and model interpretation. The proposed work presents a hybrid feature selection approach with a Guided Regularized Random Forest (GRRF) classification model to assist with the diagnosis of UTI. Further the UTI prevention and course of action must be optimized via enhanced research and progressive efforts to overcome resistance from antibiotics.

1.1. Research Objective

The primary objective of this study is centered on an advanced and emerging field of research, where machine learning algorithms are applied in IoT-Fog environments to predict Urinary Tract Infections (UTIs). To achieve this, a hybrid approach for feature selection is employed, combining the Lasso CV and k-best algorithms. Moreover, for the classification of UTIs, the study utilizes the Guided Regularized Random Forest algorithm.

By integrating these innovative approaches, the study can presents an optimal solution for UTI prediction and can be used in smart toilet system.

2. Literature Review

Various research work shows the attempt to efficiently diagnosing and predicting the urinary tract infections using technical advancements. This section briefs the studies related to anticipation, analysis and prediction of UTIs.
The study (Bhatia, 2019) introduced an innovative framework that leverages the Internet of Things (IoT) to monitor, diagnose, and predict urine infection within a home-centric environment. The framework is designed with multiple layers: perception layer, analysis layer, extraction layer, prediction layer, and visualization layer, all contributing to the urine infection diagnosis process. The prediction of urine infection is achieved using a t-ANN (temporal artificial neural network) model, demonstrating an impressive accuracy rate of 93.69%.

The work (Wojno, 2020) assessed the Multiplex PCR-based genetic testing with traditional urine culture. The comparison shows that the identification of bacteria-based complications in patients with symptoms of PCR based diagnosis significantly improves the accuracy levels and speed.

The proposed methodology (Kamei, 2023) emphasizes the significance of diagnostic methods and various UTI patient features with LUTD (Lower Urinary Tract Dysfunction). The work suggests intermittent catheterization for the patients having neurogenic LUTD.

The goal of the research proposed (Leung, 2019) gives a review on the assessment, treatment, and care of pediatrics infections. The study exhibits that the bacteria Escherichia coli is 80-90 percent responsible for urine infection among children. The research recommends that prophylaxis is able to protect long term illness and recurrent UTIs.

The investigation (Gajdacs, 2020) aims to comprehensively assess the developments in resistance and the incidence of Gram-positive cocci in UTI among both patients treated at Clinical Centre over a period of ten years. It shows that Antimicrobial susceptibility with disk fusion and E-test results in prevalent presence of Enterococcus spp.

The study (Taylor, 2018) proposes machine learning approach for UTI diagnosis and prediction which shows the accurate and promising predictions made by XGBoost algorithm. The objective of this was to compare the models on emergency department patients suffering from urine infection.

The work (Mingkuan, 2023) proposes an algorithm for speedy differentiation of urosepsis among UTI patients. The study uses retrospective analysis for dataset screening and split it into 80:20 training-validating data. The results achieves 92.9% accuracy.

The author (Gadalla, 2019) discuss clinical (17) and immunological predictors (42) for females having uncomplicated urinary infection. The research reports that the cloudiness in urine was the most recommended predict UTI presence.

The research (AL-Khikani, 2020) intended to identify the prevalence of K. oxytoca in UTI individuals who were suffering from severe infections so that urology doctors could administer the proper practical antimicrobial treatment for this pathogen. The outcome represents that K. oxytoca increases the burden in UTI and changes the sensitivity towards antimicrobial agents.

The study (Bahati, 2021) aimed to ascertain the incidence of symptoms UTI, detect the microorganisms, and assess their vulnerability to various antimicrobial medications in pregnant women. The work involves 400 pregnant women having symptomatic UTI. It clearly shows that Klebsiella pneumoniae is the most prominent followed with E.coli.

The work (Ozkan, 2018) demonstrates the use of ANN in detecting and predicting UTI and it achieves the highest level of accuracy i.e 98.3% in comparison with the other existing models such as SVM, RF and DT. The dataset comprises of 59 patients where females are 35 and males are 24.

The author (Oliveira, 2020) suggests a thorough analysis of the pathogenesis, clinical signs, imaging test, diagnosis, treatment, chemoprophylaxis, and effects of urinary tract infection in pediatric patients. The findings of the research denotes that prophylaxis is able to protect long term illness and recurrent UTIs.

The study (Homeyer, 2019), a liquid-infused nitric-oxide-releasing (LINOREl) urinary catheter was created by adding silicone oil and the nitric oxide (NO) donor S-nitroso-N-acetylpenicillamine (SNAP). The integration of the non-fouling qualities of materials with liquid injected into them, this synergistic combination enhances NO-releasing materials by minimizing SNAP leaking and enhancing the release of NO. This combination aids in the prevention of CAUTI.

The study (Leticia-Kriegel, 2019) examined to investigate the evolution of CAUTI risk. In addition, to determine whether risk variables such as age, sex, patient type (surgical vs. medical), and comorbidities affected how long it took from catheter placement to a CAUTI occurrence.

The study (Lerner, 2021) suggests proliferating of histologically, benign prostatic hyperplasia (BPH) is defined as the presence of smooth muscle and epithelial cells in the prostatic transition zone. Lower urinary tract symptoms (LUTS) that are widespread and worsen with age have an impact on society health and wellbeing.

The Study (Chotiprasitsakul, 2021) provides Clinical signs and uropathogenic detection is used to identify urinary tract infections. Antibiotics are often not recommended in cases with candiduria and urine with no growth. In order to explain the distribution of microorganisms in urine and to differentiate between bacteriuria, candiduria, and no-growth urine, we set out to build a prediction score.

The approach (Gupta, 2023) uses the IoT-fog computing based framework in which XGBoost algorithm detects and predict UTI. The model developed gives the accuracy rate of 91.45% that represents the promising improvement as compared with the other baseline techniques.
The paper (Jamaluddin, 2020) represents KNN approach to detect UTI. The study shows that at the value of k=6, the algorithm give prominent accuracy level of 97.4%. This indicates that the developed application has the ability to classify UTI correctly.

3. Proposed Methodology

The symptoms and warning signs of a urinary tract infection (UTI) can vary depending on which part of the urinary system is affected. Diagnosing a UTI typically involves medical professionals taking a urine sample for urinalysis and a urine culture. A urine culture is particularly useful for identifying the specific bacteria responsible for the infection and determining the most effective antibiotic treatment. Urinalysis, on the other hand, helps in assessing the urine for the presence of bacteria, white blood cells, and red blood cells. The outcomes of standard tests and definitive diagnosis for UTI patients were compiled to build a UTI dataset. The stages of dimensionality reduction and feature selection in data analysis and modeling for the investigation of UTIs are crucial. To increase the precision and potency of prediction models or analyses, these techniques include locating and choosing pertinent elements from the data at hand. In this research, a hybrid feature selection approach was employed, which combines the k-best and Lasso CV methods with the Guided Regularized Random Forest (GRRF) classification model to identify and predict UTIs. Figure 1 illustrates the proposed methodology.

![Proposed Framework](image)

**Figure 1 Proposed Framework.**

3.1. Data Acquisition and Data Preprocessing

The study conducted retrospective analysis on adult emergency department visits focusing on urine culture findings at a single center and multiple sites. To protect patient privacy, data were de-identified after the first database access and before analysis. De-identified data was exclusively used for the analyses. Predictive variables were sourced from the time of the emergency department (ED) visit to admission or discharge. Variables such as vital signs, lab results, urinalysis, urine dipstick results, outpatient medications, medical history, primary complaint, physical exam findings, and demographic information were employed for prediction purposes. Notably, medications taken during the ED visit and the ED diagnosis were excluded from the variables to reduce the impact of provider expertise on the prediction model. The sample dataset for urinary tract infections (UTIs) contained 1117 rows and 17 columns (Taylor, 2018).

The normalization of the data is essential to preserve the integrity of the link that exists between the variables, the results that are obtained from the analysis, and the functionality of the network. The method of normalization involves scaling each piece of data included in the dataset between the maximum and minimum values specified by the activation function. During analysis and modelling, normalization ensures that features of varying scales are treated equally. When preparing UTI data, normalization is used to scale any numerical characteristics that need it. A "sigmoid" activation function was used. In the course of the information were normalized for execution so that they fell within the bounds. [0,1] by making use of equation 1.

\[
\text{normalized\_value} = \frac{(x - \text{min\_value})}{(\text{max\_value} - \text{min\_value})}
\]

(1)

Where, \(x\) = original value, \(\text{min\_value}\) = dataset’s feature’s minimum value, \(\text{max\_value}\) = dataset’s feature’s maximum value.
3.2. Dimensionality Reduction using PCA

Principal Component Analysis (PCA) is a widely used data mining technique for reducing the dimensionality of large datasets while preserving the most significant aspects of the data’s variability. PCA identifies the principal components, which are linear combinations of the original variables that capture the highest variance in the data. These components are orthogonal to each other, meaning they are uncorrelated, and they are ordered by the amount of variability they account for. The primary advantage of PCA is its ability to filter out background noise while highlighting essential patterns in the data, making it an effective method for dimensionality reduction. By using a relatively small number of principal components that capture more than 80% of the original variables’ information, researchers can simplify complex multivariate analyses. Standardizing raw data is crucial to remove the influence of different dimensions in PCA.

Assuming n objects, let $x_{i1}, x_{i2}, ..., x_{ip}$ represent the p indexes of the ith item. Use the following matrix (2) to represent all observations of p indices of n objects.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ x_{21} & x_{22} & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{np} \end{bmatrix}$$ (2)

Where n and p = Number of objects and variables

Standardization operations on the p indexes of the n items should be repeated on the basis of the following equation (3).

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}; \quad i = 1, 2, ..., n; \quad k = 1, 2, ..., p$$ (3)

Where, $\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$, $s_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}$

The correlation coefficient covariance matrix of standard indicators may be computed by using the usual equation for determining Pearson’s correlation coefficient. The correlation coefficient may be calculated using the following equation (4):

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x-y)(y-y)}{\sqrt{\sum (x-y)^2} \sqrt{\sum (y-y)^2}}$$ (4)

Utilizing PCA with R’s correlation matrix, it is possible to compute various essential components’ Eigenvalues, eigenvectors, their contribution to variance, and cumulative contributions. The subsequent step involves constructing a model for evaluating the value of an index and subsequently performing a comprehensive assessment and analysis, including ranking. This ranking is based on an aggregated score obtained by combining the index’s score and its contribution rate.

3.3. Feature selection using K-best and Lasso Cv

Let’s have a look at a dataset represent by $[Z, X]$ to get a formulation of the wrapper-based k-Best Feature Selection technique. z is the m × n data matrix with p features and n occurrences, and X is the response vector of length m. Keep in mind that X is not the same as X, which previously described the decrease function’s noisy readings. Let the feature set $x := \{z_1, z_2, ..., z_p\}$ be the case, with $z_i$ being the ith feature in $Z' \subset Z, L_c(Z', X)$ is the real data set's measurement of a specific performance criterion in relation to a wrappers predictord, with respect to a k-dimensional subset $Z' Z$. The anticipated classification accuracy used as an example was obtained using set of data was collected from all of the participants using a 5-fold cross-validation technique. Due to the lack of universal knowledge on $L_c$, the infomation in the dataset is used for developing a classifier and measures the value of $f(x, Z', X)$, where $X_c = L_c + \epsilon$. The non-empty feature set $Z^*$, which is specified as the wrapper-based Feature Selection issue in equation (5), must be present.

<table>
<thead>
<tr>
<th>Pat_id</th>
<th>UCX_abnormal</th>
<th>ua_bili</th>
<th>ua_blood</th>
<th>ua_color</th>
<th>ua_clarity</th>
<th>ua_ph</th>
<th>ua_glucose</th>
<th>ua_bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>negative</td>
<td>moderate</td>
<td>yellow</td>
<td>clear</td>
<td>7.5</td>
<td>negative</td>
<td>few</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Small</td>
<td>negative</td>
<td>Pale yellow</td>
<td>clear</td>
<td>5</td>
<td>negative</td>
<td>many</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>negative</td>
<td>negative</td>
<td>yellow</td>
<td>clear</td>
<td>5</td>
<td>negative</td>
<td>moderate</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>negative</td>
<td>negative</td>
<td>red</td>
<td>Not clear</td>
<td>5</td>
<td>negative</td>
<td>Not reported</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>negative</td>
<td>negative</td>
<td>orange</td>
<td>clear</td>
<td>6</td>
<td>small</td>
<td>few</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>negative</td>
<td>negative</td>
<td>yellow</td>
<td>Not clear</td>
<td>6</td>
<td>negative</td>
<td>many</td>
</tr>
<tr>
<td>7</td>
<td>No</td>
<td>Small</td>
<td>Small</td>
<td>yellow</td>
<td>Not clear</td>
<td>6</td>
<td>negative</td>
<td>moderate</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Large</td>
<td>Small</td>
<td>yellow</td>
<td>clear</td>
<td>6</td>
<td>small</td>
<td>Not marked</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>moderate</td>
<td>Large</td>
<td>yellow</td>
<td>Not clear</td>
<td>5.6</td>
<td>negative</td>
<td>few</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>moderate</td>
<td>moderate</td>
<td>yellow</td>
<td>clear</td>
<td>6.5</td>
<td>negative</td>
<td>few</td>
</tr>
</tbody>
</table>

Table 1 Sample dataset.
The following is how the data was processed. The first step was to normalize all radiomic characteristics using the StandardScaler function by subtracting the mean and dividing by the standard deviation, and each set of feature values was then transformed into a mean of 0 with a variance of 1. The best parameter was then discovered by minimizing the average mean square error following a 10-fold cross-validation based on standardized features. Following the calculation of the coefficients for each feature using the Lasso function, the pertinent features were chosen based on the best parameters, and radiomic features with non-zero coefficients were found.

### 3.4. Classification using Guided Regularized Random Forest (GRRF) algorithm

In ensemble classification techniques, rather than using the findings of a single classifier, the results generated by a number of different classifiers are employed. The RF approach is one of the examples for ensemble classification that receives the greatest attention and use. Classifiers of the RF kind are characterized by their usage of randomly generated data gleaned from real-world scenarios and their composition of several trees. In order to correctly categories a sample, an input vector is assigned to each tree in the forest, and then a result is generated for each tree individually. As the conclusion, the RF algorithm picks the category that received the highest support from the audience.

The ideal of each the node's chosen at random attributes are used by RF in order to create branches off of each node. The basic operating concept of the random forest is shown in figure 2. This ensures that every variable is considered. When using the RF approach, using chosen bootstrap samples, structures are built, and the distance between nodes is achieved via the use of randomly chosen n estimators. It is important to point out that the overall amount of estimations is far more than the total number of n units. Every tree of choices has not been trimmed but has been preserved in its fullest form. In Every leaf node in trees of classification is intended for holding only one the individuals that belong to the same class. The RF approach creates more accurate generalizations and reliable estimations because it incorporates enhanced random sampling and other ensemble method techniques. Due to the lower bias findings and low correlation across trees, the RF technique has a higher estimate confidence. Because extremely huge trees were made, there was very little bias found. To streamline the classification process, we employ GRRF to narrow down the characteristics that will be used. Each feature $x_k$ GRRF improvement is shown as equation (6):

\[
G_{GRRF}(x_k,v) = \begin{cases} 
G(x_k,v) & \text{if } k \notin F \\
\lambda G(x_k,v) & \text{if } k \in F
\end{cases}
\]

(6)

F is the subset of features that were chosen to be utilized in the preceding node's instance splitting, and $\lambda \in [0,1]$ is a penalized percentage for the attributes that were not chosen. Because features with a gain value of zero aren’t considered in the selection process, GRRF is able to choose features that are not redundant.

GRRF (Guided Regularized Random Forest) presents a novel method that combines the advantages of Random Forest and regularization techniques to tackle feature selection and model interpretation challenges in healthcare applications.

---

**Figure 2** Fundamental of RF operation.
4. Results and discussion

In this section, we present the results obtained from the implementation of this study. It encompasses the dataset description, standard performance metrics, experimental outcomes, and a comprehensive analysis of performance and comparisons.

4.1. Dataset Description

The dataset (Taylor, 2018) used in this study comprises predictor variables encompassing laboratory results, urine dipstick results, urinalysis, past medical history, structural historical findings, physical exam findings, chief complaints, and demographic information. The dataset exists in two versions: one with a reduced set of 10 variables and another with a full set of 211 variables. It was obtained from an IoT-based fog environment and can be accessed through the link: https://figshare.com/articles/dataset/Predicting_urinary_tract_infections_in_the_emergency_department_with_machine_learning/5959417?file=10.

The dataset utilized in this research comprises 80,387 attributes and 219 rows, and some of the key attributes are listed below:
- Age: This attribute represents the patient's age.
- Sex: This attribute denotes the patient's gender.
- Diabetes: The presence of diabetes is indicated using a binary value (1 for presence, 0 for absence).
- Hypertension: The existence of hypertension is represented by a binary variable (1 for presence, 0 for absence).
- UTI history: This attribute indicates the presence or absence of a history of Urinary Tract Infections (UTI).
- Fever: A binary attribute indicating whether the patient has a fever (1 for presence, 0 for absence).
- Dysuria: A binary attribute indicating whether the patient experiences discomfort or pain during urination (1 for presence, 0 for absence).
- Urgency: A binary attribute indicating whether the patient feels the need to urinate urgently (1 for presence, 0 for absence).

These attributes play a significant role in the prediction of urinary tract infections within the research context.

4.2. Performance metrics

Performance metrics have become an integral part of each machine learning model. The following metrics, namely precision, accuracy, f1-score, specificity, recall, and sensitivity, are commonly utilized to thoroughly analyze the classification model. These metrics provide valuable insights into the model's effectiveness and its ability to correctly classify instances across different classes.

4.3. Experimental results and comparative analysis

The significance of the hybrid approach for feature selection is evaluated on the basis of the execution time and accuracy. The table 2 and figure 3 shows the execution time and accuracy of with or without hybrid feature selection method on the UTI dataset.

Now, analyze the outcomes that were obtained by putting the suggested technique into practice. Given the complexity of the symptoms, the purpose of this research is to develop hybrid feature selection approach with Guided Regularized Random Forest (GRRF) classification model that will facilitate the diagnosis of a urinary tract infection (UTI). Comparisons are made between the approach that has been presented and other methods that already exist, such as the ANN (Ozkan, 2018), XGBoost (Gupta, 2023) and the k-Nearest Neighbors (Jamaluddin, 2020) method. Accuracy, precision, recall, specificity, mean absolute error (MAE), and mean squared error (MSE) are performance measurements for the proposed methodology. Table 3 shows the performance measurements of the proposed methodology.

<table>
<thead>
<tr>
<th>Table 2 Proposed Method Evaluation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without hybrid feature selection -Method 1</td>
</tr>
<tr>
<td>Execution Time</td>
</tr>
<tr>
<td>0.16 (s)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3 Performance measurements of proposed method.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td>Proposed Method</td>
</tr>
</tbody>
</table>

The figure 4 represents the graphical view of the performance measurements obtained by the proposed methodology.
The dataset used in the proposed study was collected from an IoT-based fog environment. Therefore, for the comparative analysis, other studies that utilize similar types of datasets from IoT-based fog environments have been taken into consideration. By doing so, a relevant and meaningful comparison can be made with studies that share similar data characteristics and settings. The performance of the proposed model is compared with existing state of art in terms of accuracy, precision, recall and specificity. The comparative view is shown in the table 4 and figure 5.

### Table 4 Comparative review.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN (Ozkan, 2018)</td>
<td>98.30</td>
<td>100</td>
<td>97.77</td>
<td>100</td>
</tr>
<tr>
<td>XGBoost(Gupta, 2023)</td>
<td>91.45</td>
<td>95.49</td>
<td>84.79</td>
<td>95.96</td>
</tr>
<tr>
<td>KNN(Jamaluddin, 2020)</td>
<td>97.40</td>
<td>94.68</td>
<td>94.58</td>
<td>92.58</td>
</tr>
<tr>
<td>Proposed work</td>
<td>98.88</td>
<td>98.90</td>
<td>98.72</td>
<td>97.45</td>
</tr>
</tbody>
</table>
4.3.1. Accuracy

Accuracy is a frequently used parameter to assess a forecasting model’s effectiveness. It calculates the percentage of the system’s overall forecasts that were accurate predictions. Figure 5 displays the accuracy for stated and proposed methods. The suggested procedure is having more accuracy of 98.8% when compared with the current ones.

\[
A = \frac{T_{pos} + T_{neg}}{T_{pos} + T_{neg} + F_{pos} + F_{neg}}
\]

\(T_{pos}=\)true positive
\(T_{neg}=\)true negative
\(F_{pos}=\)false positive
\(F_{neg}=\)false negative

4.3.2. Precision

Regarding all expected positive situations, precision focuses on the percentage of correctly forecast positive instances. It helps evaluate the model’s ability to cut down on erroneous positives. It provides the percentage of information points that have been labeled as infected that are actually affected. Figure 4 displays the precision of both the suggested and existing approaches. The suggested procedure achieves 94.68% precision. However the precision value (Ozkan, 2018) is remarkable with 100%.

The mathematical expression is as follows:

\[
P = \frac{T_{pos}}{T_{pos} + F_{pos}}
\]

4.3.3. Recall

Recall, also referred to as sensitivity or the true affirmative rate, evaluates the ability of a model to correctly identify positive cases among all actual positive outcomes. It is a key metric for assessing a model’s effectiveness in reducing false negatives, measuring the accuracy of positive test identification. Recall quantifies how well the algorithm can identify positive samples, and it increases as more positive samples are correctly identified.

In Figure 5, the recall rates of both the proposed and current methods are displayed. Notably, the proposed methods demonstrate superior recall performance compared to state-of-the-art methods.

\[
R = \frac{T_{pos}}{T_{pos} + F_{neg}}
\]

4.3.4. Specificity

Specificity is a crucial parameter for evaluating binary classification models, measuring their ability to distinguish true negatives from false negatives. High specificity indicates fewer false positives, which is vital in scenarios like medical diagnostics. For instance, a high specificity is preferred in medical diagnostics to reduce the possibility of incorrectly categorizing healthy people as having a condition. These measures add up to a more thorough assessment of the model's
effectiveness. The specificity rates are depicted in Figure 5. The proposed method achieves 97.45 whereas the method (Ozkan, 2018) achieves 100%.

\[
\text{Specificity} = \frac{T_{\text{neg}}}{P_{\text{pos}} + T_{\text{neg}}} \tag{10}
\]

The proposed system's ROC Curve is shown in Figure 6. When contrasted to the current methods, the one being suggested has a greater rate. The efficacy of a binary categorization system as its ability to discriminate threshold is changed illustrated graphically by the ROC curve. However, because the ROC curve is normally used to evaluate model classification, it cannot be utilized to analyse multi-modal image fusion directly.

![ROC Curve](image)

**Figure 6 ROC Curve.**

### 4.4. Accuracy and Loss for Training and Validation

The accuracy of a urinary tract infection is illustrated in Figure 7, and it refers to how well it can efficiently integrate and preserve pertinent information from the input pictures while eliminating noise, antiques, and irregularities.

![Training and Testing Accuracy](image)

**Figure 7 Training and Testing Accuracy.**
The difference between the fused image's quality or fidelity and the original input photos is shown in Figure 8. It shows the degree in which the process of fusion generates artifacts or discrepancies, fails to safeguard important information, or otherwise affects the quality.

![Figure 8 Model loss for Training and validation.](image)

The increase in error rates, particularly Mean Absolute Error (MAE) and Mean Squared Error (MSE), is a crucial aspect to understand in evaluating the performance of predictive models in healthcare, such as the one used for UTI prediction. Several factors contribute to these errors, including data quality issues like inaccuracies and variations in data collection, the complexity of the predictive model, sample size limitations, feature selection, model tuning, data drift over time, and unmodeled variables that impact UTI but aren't considered in the model. Recognizing these factors and their potential influence on error rates is essential for a comprehensive understanding of the model's limitations and areas for further research and improvement. This transparency is vital for the responsible and informed use of predictive models in healthcare applications.

5. Conclusions

In conclusion, the hybrid feature selection technique for UTI detection and prediction offers a promising opportunity to enhance the accuracy and effectiveness of UTI diagnostics. By combining multiple feature selection methods, it efficiently identifies the most relevant features in large datasets, ensuring that they are biologically meaningful and shed light on the root causes of UTIs. Moreover, this approach minimizes the limitations and biases associated with using a single method, increasing its robustness. The Guided Regularized Random Forest (GRRF) method presents a potential strategy for UTI detection and prediction by amalgamating the strengths of Random Forest and regularization techniques, further improving feature selection.

As UTIs are prevalent in modern society, future research and efforts are essential to optimize prevention and treatment methods while combating antibiotic resistance. With a remarkable 98.9% precision rate, 98.88% accuracy, 98.72% recall rate, 97.45% specificity, 18% mean absolute error, and 32% mean squared error, this study's high predictive accuracy supports its practical application in real-time settings. Additionally, further enhancements could be achieved by incorporating other hybrid machine learning algorithms, making the system suitable for deployment in various environments, from public restrooms to hospitals, thus improving UTI diagnosis across diverse scenarios.

The study's high prediction rate enables its practical application in real-time scenarios. In the future, the incorporation of various other hybrid machine learning algorithms could further enhance the prediction rate. Additionally, the proposed system holds potential for deployment in real-world settings, including public restrooms, hospitals, and offices, thereby increasing the effectiveness and feasibility of diagnosing UTIs in diverse environments.

5.1. Limitation

While the proposed system demonstrates the capability to achieve optimal accuracy, it is not without limitations and challenges. One significant limitation is its heavy reliance on the availability and quality of data, which can pose challenges in obtaining high-quality UTI-related features. Additionally, the implementation of smart toilets equipped with IoT sensors presents significant challenges and incurs higher costs, further adding to the complexities of the system.
Ethical considerations

Not applicable.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This research did not receive any grant from external funding agencies.

References


Behzadi, P., Behzadi, E., & Pawlak-Adamska, E. A. (2019). Urinary tract infections (UTIs) or genital tract infections (GTIs)? It’s the diagnostics that count. GMS hygiene and infection control, 14.


