A Hybrid Feature Selection Approach for Urinary Tract Infection Detection and Prediction in IoT-Fog Environment

Anisha[1], Munish Sabharwal[2], Rohit Tripathi[3]
[1,2]Galgotias University, Greater Noida, India
[3]JC Bose University of Science & Technology, YMCA, Faridabad, India

[1]anishanagpal@outlook.com, [2]mscheckmail@yahoo.com, [3]rohittripathi30.iitd@gmail.com

Abstract

Urinary Tract Infections (UTIs) are a common health issue that millions of people experience globally and a significant impact on the general health. Depending on the location and intensity of the infection, UTIs appear differently clinically. Dysuria, frequency, urgency, suprapubic discomfort, and hematuria are typical symptoms. Patients with severe conditions could have a fever, flank discomfort, and other systemic symptoms that point to an upper UTI. Clinical assessment and lab tests are used to get the precise diagnosis of UTIs. The primary aim of this research is to utilize appropriate Machine Learning (ML)-based algorithms to predict Urinary Tract Infections (UTIs) in IoT-Fog environments. The ultimate goal is to develop a predictive model that can be effectively implemented in a smart toilet system. By achieving this objective, the study aims to offer an innovative and practical solution for UTI prediction, leveraging the potential of ML algorithms in IoT-Fog environments to enhance healthcare and improve public health. This paper presents hybrid approach for feature selection and using Guided Regularized Random Forest (GRRF) classification to assist with the diagnosis of UTI. Data from regular exams and definitive diagnostic results for UTI patients were used to generate a UTI dataset. Principle Component Analysis (PCA) is used for dimensionality reduction, while K-best and Lasso CV are used for feature selection. Using our suggested strategy, this research was able to identify UTI with a 98.8% accuracy and 98.90% precision rate. 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. [1,2].

In the future, fog computing is expected to play a pivotal role in meeting the escalating demand for real-time 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[3].

Machine learning predictive models 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. The convergence of IoT and ML contributes to improving healthcare services and provides valuable insights into proactive and personalized patient care for better health outcomes [4,5].

The goal of the present study is to develop a Machine Learning (ML)-based algorithm for predicting Urinary Tract Infections (UTIs) 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 [6].

Numerous people across the world suffer from urinary tract infections (UTIs), which are a prevalent medical problem [7]. UTIs are brought on by bacteria that enter and attack the urinary tract, which consists of the bladder, urethra, ureters, and kidneys [8]. 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. [9]. 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 [10]. 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 practicing excellent hygiene, consuming lots of water, urinating regularly, and clearing the bladder both before and after sexual activity [11]. 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 [12]. 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 [13].

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 [14].

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 [15]. 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 [16]. It can get over the drawbacks associated with separate procedures and provide a more thorough evaluation of the feature space by integrating various approaches [17].

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 Novelty

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.

1.2 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.

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.

Bhatia [33] 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%.

Kirk J. Wojno[19] 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 with PCR based diagnosis significantly improves the accuracy levels and speed.

Jun Kamei [20] 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 by Leung [21] is to give 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 the regular antimicrobial prophylaxis in children with acute UTI symptoms.

The investigation of Márió Gajdács[22] 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.

Taylor[23] 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.

Mingkuan Su[24] proposes 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.

Gadalla[5] discuss clinical (17) and immunological predictors (42) for females having uncomplicated urine infection. The research reports that the cloudiness in urine was the most recommended predict UTI presence.

The research of AL-Khikani[25] 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 of Bahati Johnson [26] 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.

Ozkan [27] 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.

Oliveira[28]suggests a thorough analysis of the pathogenesis, clinical signs, imaging test, diagnosis, treatment, chemo-prophylaxis, 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[29], 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 leaching and enhancing the release of NO. This combination aids in the prevention of CAUTI.

The study of Homeyer [30] 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 [31] 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 of Chotiprasitsakul [32] 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 proposed in [33] 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[34] 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

According to which part of the urinary system is impacted, the symptoms and warning signs of a UTI may differ. A medical professional may take a urine sample for urinalysis and a urine culture to diagnose a UTI. While a urine culture identifies the precise bacteria causing the illness and establishes the most efficient antibiotic therapy, a urine analysis aids in testing the urine for the presence of germs, 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. We used a hybrid feature selection approach using k-best and Lasso Cv with Guided Regularized Random Forest (GRRF) classification model in the research to identify and predict UTIs. The proposed methodology is depicted in figure 1.

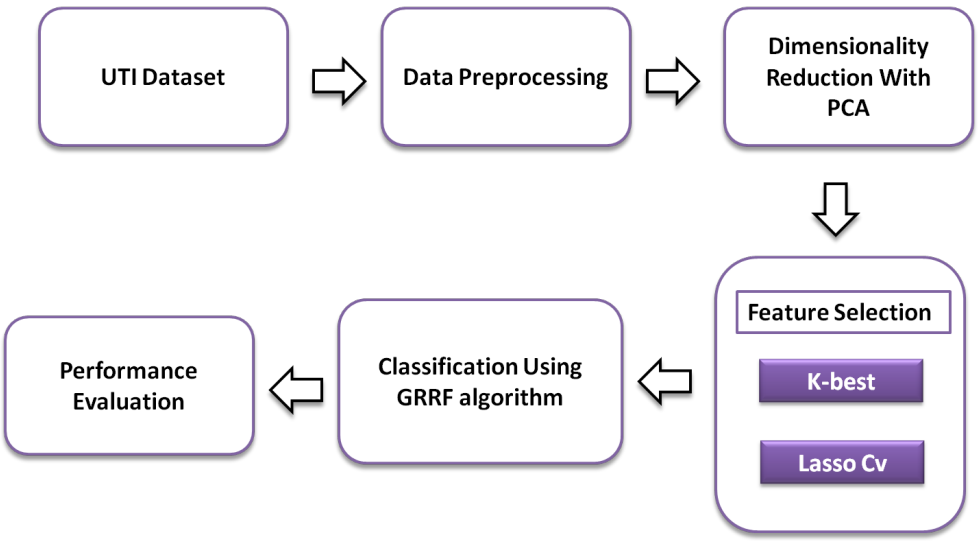

Figure 1: Proposed system with hybrid feature selection.

3.1 Dataset

Adult emergency department visits with urine culture findings were analyzed retrospectively by single centre and several sites. After the first database access but before analysis, the data were de-identified. Analyses only utilized and kept de-identified data. Each emergency department (ED) visit's individual data pieces were sourced from the business data repository. As variables for prediction, only information accessible from the time of the ED visit to admission or discharge was employed. To decrease the impact of provider expertise on the prediction model, the medications taken during the ED visit and the ED diagnosis were not included as variables. Vital signs, lab results, urinalysis and urine dipstick results, current outpatient medicines, prior medical history, the primary complaint, and structured historical and physical exam findings were some of the predictive factors. Demographic information (age, sex, ethnicity, etc.) was another factor. Table 1 shows the sample dataset for UTIs that contain 1117 rows × 17 columns [27].

Table 1: Sample dataset

Number Of Patients	Temperature of patient	Lumbar pain	Urine pushing (continuous need for urination)	Micturition pains	Burning of urethra, itch, swelling of urethra outlet	Inflammation of urinary bladder	Nephritis of renal pelvis origin
0	35.5	2	1	1	1	1	1
1	35.5	1	2	2	2	2	1
2	35.5	2	1	1	1	1	1
3	35.5	1	2	2	2	2	1
4	35.5	2	1	1	1	1	1
…	…	…	…	…	…	…	…
1112	41.5	2	2	2	2	2	2
1113	41.5	1	1	1	1	1	1
1114	41.5	1	1	1	1	1	1
1115	41.5	1	1	1	1	1	1
1116	41.5	1	1	1	1	1	1

3.2 Preprocessing

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.

$$normalized\_value = (x - min\_value) / (max\_value - min\_value) \quad (1)$$

Where, x = original value.

min_value=dataset's feature's minimum value.

max_value=dataset's feature's maximum value.

3.3 Dimensionality Reduction using PCA

For data sets with a large number of dimensions, PCA may help cut down on the number of dimensions while still accurately reflecting the data's variability. Extracting the principle components in the direction of data displaying the highest variability, PCA is a well-known data mining approach.The first one is the largest variance, and the others are orthogonal to each other under the extra restriction that they include the highest amount of variability. The primary benefit of PCA is its ability to ignore background noise while still isolating significant patterns in the data. It is also the most effective dimensionality reduction method in capturing data variability.

If you're going to do a quantitative analysis of variables, it seems to reason that you'd want to have more information to work with. The computational effort and complexity of problem analysis in the study of multi-variable situations rises as the number of variables rises. For this purpose, principal components analysis is the best tool available. Since the goal of the principle component is to reduce the dimensionality of a problem and, by extension, the number of variables, a relatively small number of principle components is often utilized in applied research as long as they can contain more than 80% of the information of the original variables. This is because reducing dimensionality is the main goal of the component itself.Raw data should be standardized to eliminate the effect of dimensions.Assuming n objects, let x_{i1} , x_{i2} ,..., and x_{ip} represent the p indexes of the i[th]item. Use the following matrix (2) to represent all observations of p indices of n objects.

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

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).

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

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):

$$r = \frac{\sigma^2{}_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \;\; \sqrt{\sum (y-\bar{y})^2}}} \quad (4)$$

Applying PCA with R's correlation matrix, we can calculate each principle component's Eigen value, eigenvector, contribution to the variance, and cumulative contribution.Develop a model for assessing the value of an index, and then conduct a complete review and analysis, including

ranking, based on a comprehensive score calculated using the index's score and the contribution rate.

3.4 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 $m \times n$ 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 $x \coloneqq \{z_1, z_2, \ldots z_p\}$be the case, with $z_I$being the i th 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 predictor d, 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 onL_c, the $d$information in the dataset is used for developing a classifier.and measures the value of $X_c(Z^{'}, X)$, where $X_c = L_c + \varepsilon$.The non-empty feature setZ^*, which is specified as the wrapper-based Feature Selection issue in equation (5), must be present.

$$Z^* \coloneqq \arg \min_{Z^{'} \subset Z} X_c(Z^{'}, X) \tag{5}$$

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.

3.5 Classification using Guided Regularized Random Forest (GRRF) algorithm

In ensemble classification techniques, rather of 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.

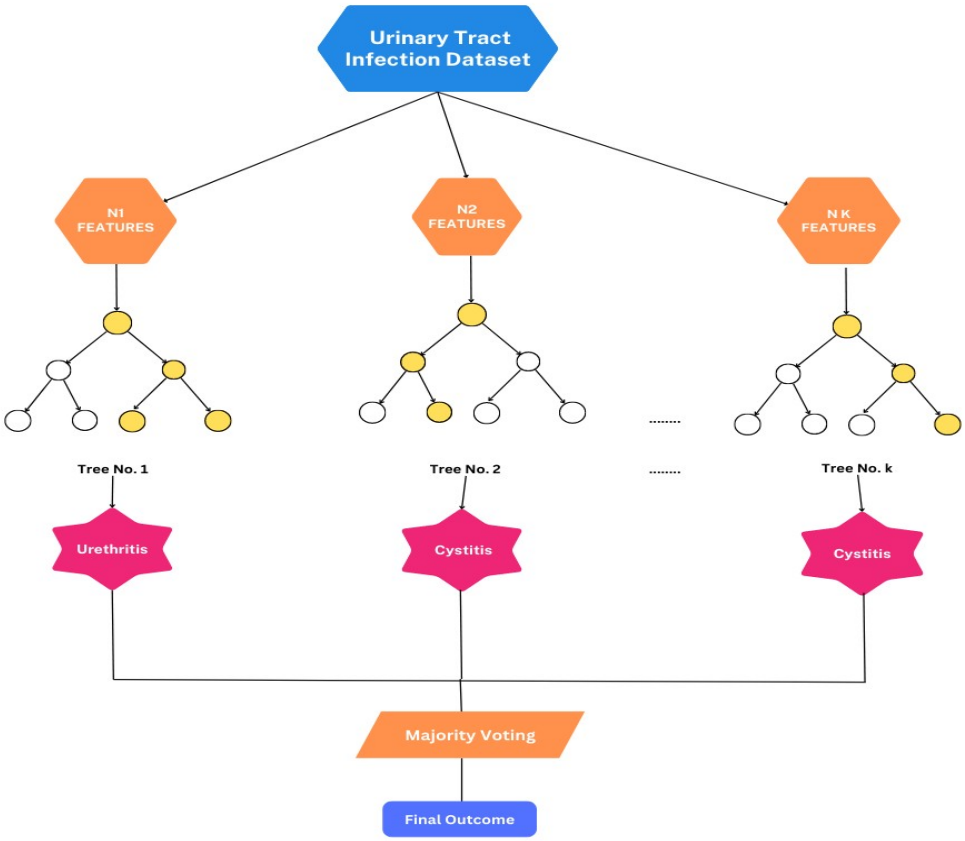

Figure 2: Fundamental of RF operation

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):

$$\left.\begin{matrix} G_{RRF}(x_k,v)=G(x_k,v) if\ k \in F \\ G_{RRF}(x_k,v)=\lambda G(x_k,v) if\ k \notin F \end{matrix}\right\} \quad (6)$$

$F$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.

4. Result 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 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 shows the execution time and accuracy of with or without hybrid feature selection method on the UTI dataset.

Table 2: Evaluation of the proposed method

Without hybrid feature selection		With hybrid feature selection	
Execution Time	**Accuracy**	**Execution Time**	**Accuracy**
0.16 (s)	98%	0.34 (s)	98.8%

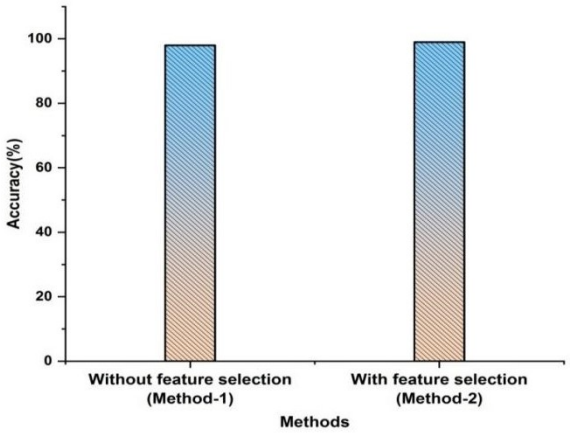

Figure 3: Accuracy of proposed method

Now, we will 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(21), XGBoost (27) and the k-Nearest Neighbors (KNN) (28) 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 2: Performance measurements of proposed method

	Accuracy	Precision	Recall	Specificity	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
Proposed Method	98.88	98.90	98.72	97.45	18	32

The figure 4 represents the graphical view of the performance measurements obtained by the proposed methodology.

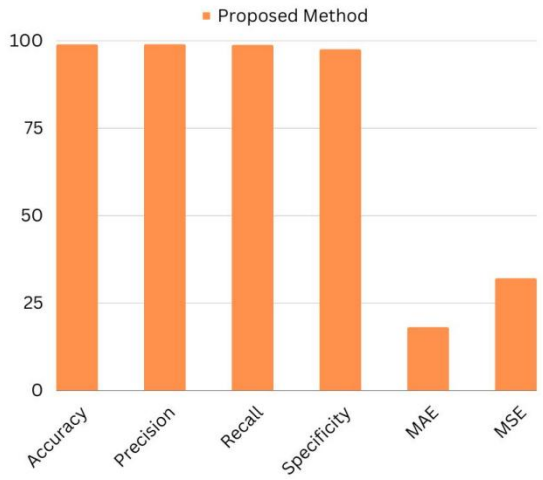

Figure 4: Performance Measurements of Proposed System

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 3.

Table 3: Comparative view

Methods	Accuracy	Precision	Recall	Specificity
ANN [27]	98.30	100	97.77	100
XGBoost[33]	91.45	95.49	84.79	95.96
KNN[34]	97.40	94.68	94.58	92.58

Proposed work	98.88	98.90	98.72	97.45

Accuracy rate

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 3 displays the accuracy for stated and proposed methods. The suggested procedures are more precise when compared with the current ones.

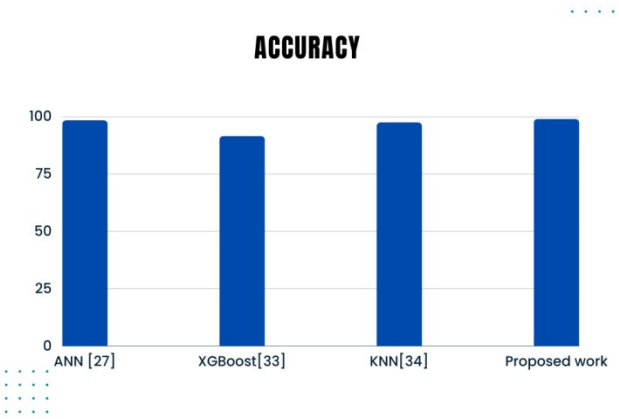

Figure 5: Comparison of Accuracy

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 [21] is remarkable with 100%.

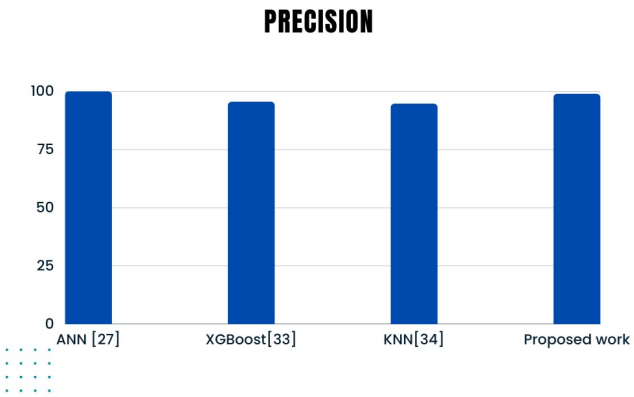

Figure 6: Comparison of Precision

Recall

Recall, also known as sensitivity or the true affirmative rate, is the proportion of appropriately predicted positive cases among all really positive outcomes. It aids in assessing the model's capacity to reduce false negatives. The proportion of samples that were correctly identified as being positive (recall) is what is used to measure the accuracy of positive test identification. The recall measures how accurately the algorithm can identify Positive samples. As additional favorable samples are found, recall rises. Figure 5 displays the recall of the suggested and current methods. The proposed methods perform better in terms of recall than cutting-edge methods.

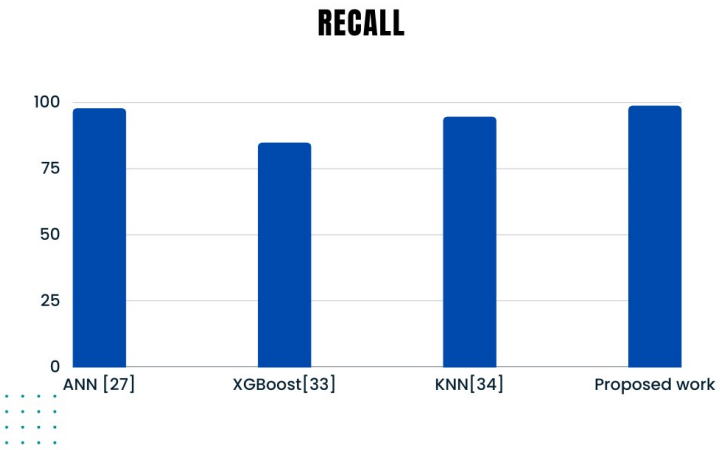

Figure 7: Comparison of Recall

Specificity

Specificity is a productivity parameter that measures how well a binary classification model can distinguish the real negatives from the rest of the false negatives. It gauges the percentage of genuine negatives that the algorithm accurately identified. A high specificity indicates that the method is effective and has a low rate of false positives at correctly recognizing negative examples. When the cost or repercussions of false positives are substantial, specificity is especially crucial. 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 6. The proposed method achieves 97.45 whereas the method[21] achieves 100%.

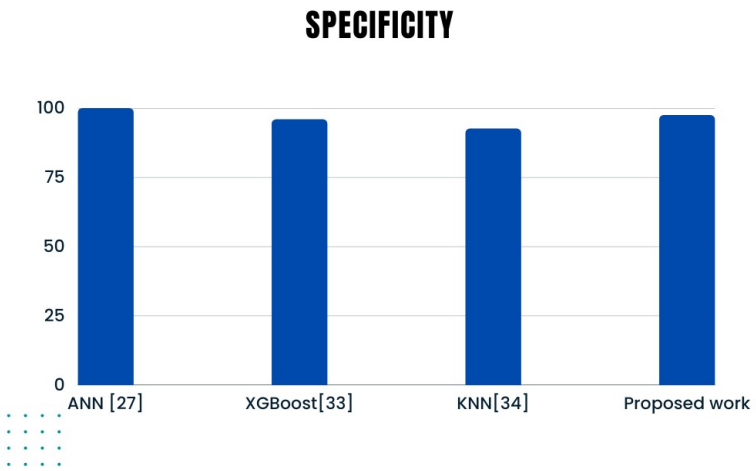

Figure 6: Comparison of Specificity

Additionally, since it defines the trade-off between true negatives and false positives, the selection of an adequate classification threshold may have an effect on the specificity rate. It is important to remember that specificity does not give a whole picture of the performance of a model. To achieve a reliable evaluation of the model's performance, it is therefore advised to take into account several metrics and evaluate the model using appropriate validation approaches, such as cross-validation.

5. Conclusion

In conclusion, the hybrid feature selection technique for UTI detection and prediction offers an exciting chance to increase the precision and effectiveness of UTI diagnostic and prediction. This strategy may successfully find the most pertinent and interesting characteristics in massive data sets by combining various feature selection strategies, including as filter, wrapper, and embedding methods. The hybrid feature selection method has a number of benefits. In the beginning, it enables a choice of variables that are most pertinent to UTI identification and forecasting by allowing for the use of domain knowledge and expert views. This guarantees that the features chosen are of biological importance and offer insightful information on the fundamental causes of UTIs. Second, combining many feature selection techniques lessens the drawbacks and biases of using just one methodology alone. A potential strategy for the identification and forecasting of urinary tract infections (UTIs) is the Guided Regularized Random Forest (GRRF) method. It offers a powerful and reliable system for enhancing the precision and comprehension of UTI prognosis and diagnosis models by combining the advantages of Randomized Forest and regularization techniques. By maximizing the benefits of many approaches and reducing the drawbacks of others, it improves the robustness of selecting features. Urinary tract infections are one of the most prevalent medical conditions in modern culture. Future UTI prevention and treatment plans must be optimized via further research and ongoing efforts to overcome antibiotic resistance. Using our proposed strategy, this research was able to identify UTI with a 98.9% precision rate, 98.88% accuracy, 98.72% recall rate, 97.45% of Specificity, 18% of Mean Absolute Error and 32% of Mean Squared Error.

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.

Reference

1. Bansal, M.; Sirpal, V.; Choudhary, M.K. Advancing e-Government using Internet of Things. In Mobile Computing and Sustainable Informatics; Springer: Berlin/Heidelberg, Germany, 2022; pp. 123–137
2. Lin, Y.W.; Lin, Y.B.; Liu, C.Y. AItalk: A tutorial to implement AI as IoT devices. IET Netw. 2019, 8, 195–202. [CrossRef]
3. Alshamrani, M. IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey. J. King Saud Univ.-Comput. Inf. Sci. 2022, 34, 4687–4701. [CrossRef]
4. Enshaeifar, S.; Zoha, A.; Skillman, S.; Markides, A.; Acton, S.T.; Elsaleh, T.; Kenny, M.; Rostill, H.; Nilforooshan, R.; Barnaghi, P. Machine learning methods for detecting urinary tract infection and analysing daily living activities in people with dementia. PLoS ONE 2019, 14, e0209909. [CrossRef] [PubMed]
5. Gadalla, A.A.H.; Friberg, I.M.; Kift-Morgan, A.; Zhang, J.; Eberl, M.; Topley, N.; Weeks, I.; Cuff, S.; Wootton, M.; Gal, M.; et al. Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms. Sci. Rep. 2019, 9, 19694. [CrossRef] [PubMed]
6. Ijaz, M.; Li, G.; Lin, L.; Cheikhrouhou, O.; Hamam, H.; Noor, A. Integration and Applications of Fog Computing and Cloud Computing Based on the Internet of Things for Provision of Healthcare Services at Home. Electronics 2021, 10, 1077. [CrossRef]
7. Bankar, N., Chandi, D.H., Patil, P. and Mahajan, G., 2021. Comparative Antibiogram of Escherichia Coli Isolated from the Urinary Tract Infection in Patients from Tertiary Care Hospital. *Journal of Pharmaceutical Research International*, *33*(35B), pp.123-128.
8. Verma, V. and Singh, Y., 2023, February. Identification and Monitoring of Urinary Organs in Women Genital System using Deep Learning Model. In 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 1-7). IEEE.
9. Wong, J.W., Xu, R.H., Ramm, O., Tucker, L.Y. and Zaritsky, E.F., 2023. Urinary Tract Infections Among Gender Diverse People Assigned Female at Birth on Testosterone. Urogynecology, 29(2), pp.295-301.
10. Silva, A., Costa, E., Freitas, A. and Almeida, A., 2022. Revisiting the frequency and antimicrobial resistance patterns of bacteria implicated in community urinary tract infections. Antibiotics, 11(6), p.768.
11. Baijwan, S. and Dhyani, A., 2023, February. Performance Analysis of an Inflammatory Process in the Bladder Cystitis Development using Machine Learning Approach. In 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS) (pp. 01-06). IEEE.
12. Aggarwal, N. and Lotfollahzadeh, S., 2022. Recurrent urinary tract infections. In StatPearls [Internet]. StatPearls Publishing.

13. Sulis, G., Sayood, S. and Gandra, S., 2022. Antimicrobial resistance in low-and middle-income countries: current status and future directions. Expert review of anti-infective therapy, 20(2), pp.147-160.
14. Bijlani, N., Nilforooshan, R. and Kouchaki, S., 2022. An unsupervised data-driven anomaly detection approach for adverse health conditions in people living with dementia: Cohort study. JMIR aging, 5(3), p.e38211.
15. Gehringer, C., Regeniter, A., Rentsch, K., Tschudin-Sutter, S., Bassetti, S. and Egli, A., 2021. Accuracy of urine flow cytometry and urine test strip in predicting relevant bacteriuria in different patient populations. BMC infectious diseases, 21(1), pp.1-8.
16. Hateet, RR (2022). Susceptibility to Antibiotics of Bacteria Causing Urinary Tract Infection in Pregnant Women Infected with COVID-19. Kesmas: Jurnal Kesehatan Masyaraka Nasional (National Public Health Journal) , 17 (sp1).
17. Behzadi, P., Behzadi, E. and 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.
18. Bhatia, M.; Kaur, S.; Sood, S.K. IoT-inspired smart toilet system for home-based urine infection prediction. ACM Trans. Comput. Healthc. 2020, 1, 1–25. [CrossRef]
19. Wojno, K.J., Baunoch, D., Luke, N., Opel, M., Korman, H., Kelly, C., Jafri, S.M.A., Keating, P., Hazelton, D., Hindu, S. and Makhloouf, B., 2020. Multiplex PCR based urinary tract infection (UTI) analysis compared to traditional urine culture in identifying significant pathogens in symptomatic patients. Urology, 136, pp.119-126.
20. Kamei, J., & Fujimura, T. (2023). Urinary tract infection in patients with lower urinary tract dysfunction. Journal of Infection and Chemotherapy.
21. Leung, A.K., Wong, A.H., Leung, A.A. and Hon, K.L., 2019. Urinary tract infection in children. Recent patents on inflammation & allergy drug discovery, 13(1), pp.2-18.
22. Gajdács, M., Ábrók, M., Lázár, A. and Burián, K., 2020. Increasing relevance of Gram-positive cocci in urinary tract infections: a 10-year analysis of their prevalence and resistance trends. Scientific reports, 10(1), p.17658.
23. Taylor, R. A., Moore, C. L., Cheung, K. H., & Brandt, C. (2018). Predicting urinary tract infections in the emergency department with machine learning. PloS one, 13(3), e0194085.
24. Su, M., Guo, J., Chen, H., & Huang, J. (2023). Developing a machine learning prediction algorithm for early differentiation of urosepsis from urinary tract infection. Clinical Chemistry and Laboratory Medicine (CCLM), 61(3), 521-529.
25. AL-Khikani, F. H. O., Abadi, R. M., & Ayit, A. S. (2020). Emerging carbapenemase Klebsiella oxytoca with multidrug resistance implicated in urinary tract infection. Biomedical and Biotechnology Research Journal (BBRJ), 4(2), 148-151.
26. Johnson, B., Stephen, B.M., Joseph, N., Asiphas, O., Musa, K. and Taseera, K., 2021. Prevalence and bacteriology of culture-positive urinary tract infection among pregnant women with suspected urinary tract infection at Mbarara regional referral hospital, South-Western Uganda. BMC pregnancy and childbirth, 21(1), pp.1-9.
27. Ozkan, I.A., Koklu, M. and Sert, I.U., 2018. Diagnosis of urinary tract infection based on artificial intelligence methods. Computer methods and programs in biomedicine, 166, pp.51-59.
28. Oliveira, E.A. and Mak, R.H., 2020. Urinary tract infection in pediatrics: an overview. Jornal de pediatria, 96, pp.65-79.

29. Homeyer, K.H., Goudie, M.J., Singha, P. and Handa, H., 2019. Liquid-infused nitric-oxide-releasing silicone foley urinary catheters for prevention of catheter-associated urinary tract infections. ACS biomaterials science & engineering, 5(4), pp.2021-2029.
30. Letica-Kriegel, A.S., Salmasian, H., Vawdrey, D.K., Youngerman, B.E., Green, R.A., Furuya, E.Y., Calfee, D.P. and Perotte, R., 2019. Identifying the risk factors for catheter-associated urinary tract infections: a large cross-sectional study of six hospitals. BMJ open, 9(2), p.e022137.
31. Lerner, L.B., McVary, K.T., Barry, M.J., Bixler, B.R., Dahm, P., Das, A.K., Gandhi, M.C., Kaplan, S.A., Kohler, T.S., Martin, L. and Parsons, J.K., 2021. Management of lower urinary tract symptoms attributed to benign prostatic hyperplasia: AUA guideline part I—initial work-up and medical management. The Journal of urology, 206(4), pp.806-817.
32. Chotiprasitsakul, D., Kijnithikul, A., Uamkhayan, A. and Santanirand, P., 2021. Predictive value of urinalysis and recent antibiotic exposure to distinguish between bacteriuria, Candiduria, and no-growth urine. Infection and Drug Resistance, pp.5699-5709.
33. Gupta, A., Singh, A. Prediction Framework on Early Urine Infection in IoT–Fog Environment Using XGBoost Ensemble Model. Wireless Pers Commun 131, 1013–1031 (2023).
34. Jamaluddin, M.N.F., Malik, S.N.A., Fauzi, S.S.M., Razak, T.R., Halim, I.H.A., Mohammed, A.H. and Gining, R.A.J., 2020. An Application of presumptive diagnosis for urinary tract infection via kNN algorithm approach. In Charting the Sustainable Future of ASEAN in Science and Technology: Proceedings from the 3rd International Conference on the Future of ASEAN (ICoFA) 2019-Volume 2 (pp. 377-388). Springer Singapore.