Early Urine Infection Prediction Framework using XGBoost Ensemble Model in IoT-Fog Environment

Aditya Gupta (adityag.cs.19@nitj.ac.in)
NIT Jalandhar: Dr BR Ambedkar National Institute of Technology https://orcid.org/0000-0001-8933-7221

Amritpal Singh
NIT Jalandhar: Dr BR Ambedkar National Institute of Technology

Research Article

Keywords: Urine Infection, Internet of Things, Fog Computing, XGBoost, Ensemble Learning

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1311498/v1

Noname manuscript No.
(will be inserted by the editor)

Early Urine Infection Prediction Framework using XGBoost Ensemble Model in IoT-Fog Environment

Aditya Gupta · Amritpal Singh

Received: date / Accepted: date

Abstract Urine infections are one of the most prevalent concerns for the healthcare industry that may impair the functioning of the kidney and other renal organs. As a result, early diagnosis and treatment of such infections are essential to avert any future complications. Conspicuously, in the current work, an intelligent system for the early prediction of urine infections has been presented. The proposed framework uses IoT-based sensors for data collection, followed by data encoding and Infectious Risk Factor computation using the XGBoost algorithm over the fog computing platform. Finally, the analysis results along with the health-related information of users are stored in the cloud repository for future analysis. For performance validation, extensive experiments have been carried out and results are calculated based on real-time patients' data. The statistical results of accuracy(91.45%), specificity(95.96%), sensitivity(84.79%), precision(95.49%), and f-score(90.12%) reveal the significantly improved performance of the proposed strategy over other baseline techniques.

Keywords Urine Infection · Internet of Things · Fog Computing · XGBoost · Ensemble Learning

✉ Aditya Gupta
Dr. B R Ambedkar National Institute of Technology, Jalandhar, India
E-mail: adityag.cs.19@nitj.ac.in
Amritpal Singh
Dr. B R Ambedkar National Institute of Technology, Jalandhar, India
E-mail: amritpal.singh203@gmail.com

1 Introduction

In recent years, the innovations of Industry 4.0 has uncovered a wide range of novel technologies [1]. The utilization of these technologies provide innovative and novel solutions to many problems in distinguished application domains such as agriculture, healthcare, transportation, and smart grids [2]. The adoption of data-driven technologies such as wireless sensor networks (WSN) and the Internet of Things (IoT), supported by machine learning and artificial intelligence has revolutionized the healthcare domain by providing efficient health solutions [3]. The Internet of Things includes a set of related devices with data collection and transmission capabilities through wireless media [4]. These devices generate large amounts of patient-centric health-related data. The processing of such data requires third-party cloud data centers. However, transferring a large volume of raw data to cloud servers requires a huge bandwidth. Moreover, several issues of cloud computing such as location unawareness, downtime, less security, and high latency involved make it infeasible for delay-sensitive applications. Thus, a new computing paradigm has been emerged, namely, fog computing, which serves as the backbone of delay-sensitive applications to provide users with real-time services [5].

1.1 Research Domain

Urinary infection is one of the major health problems in modern society, which may occur in any organ of the urinary system, such as the ureter, bladder, kidney, and urethra [6]. However, in most cases, this infection occurs in the bladder and urethra. Infections are caused by the growth of bacteria and fungi in the urinary tract,

leading to life-long complications. Another main reason for the prevalence of this infection among young people is improper and unhygienic diet. The basic symptoms of urinary tract infection include a burning sensation when urinating, a strong urge to urinate, heavy urine smell, and blood in the urine. Urine infections can occur in both men and women. However, compared with men, women possess a higher risk of developing such infections. According to the reports [7], almost 50 % of the women face urine infections in their lifetime. A study conducted in a nursing home in the United States [8] reported that urine infections are the most common hospital-acquired infections. If left untreated, urine infections can cause serious health problems, such as affecting the work of the kidneys and other organs. Therefore, it is vital to detect and diagnose the cause of urinary tract infections as early as possible to avoid unnecessary risks and long-term complications associated with such infections.

1.2 Motivation and Research Contributions

To detect urine infections, several urine-based tests can be performed in the pathology laboratory to measure different parameters. The traditional methods of collecting and testing direct urine infections are very cumbersome and involve multiple levels of human presence, resulting in poor accuracy of the results. Urine paper tests are used in different pathology laboratories to diagnose urine diseases. A typical urine test paper will produce 10 different color reactions when it comes in contact with urine. These tests are performed to measure the presence of parameters such as glucose, protein, hemoglobin, acetone, ketones, bilirubin, pH, and specific gravity. Table 1 shows the different diseases and effects with varying values of these parameters. A typical pathology test requires immersing the test strip in a well-mixed urine sample for a long period of time. The strip undergoes an incubation to measure the reactions. Color variation on the strip is noticed to diagnose different defects. Hence, it is clear that improper treatment at any stage may lead to the production of false results. For example, if the mixture is not properly mixed, some particles (such as white blood cells and red blood cells) will remain at the bottom of the container, or if there is too much mixture on the strip, they may distort the color. Hence, the drawbacks of traditional urine testing and the desire for early detection of urine infections suggest an urgent need for the use of the Internet of Things in this field. Different internet-enabled sensors embedded in the toilet systems are available in the market to collect urine-related information in real-time.

Hence, in our current work, we propose a smart health monitoring framework that collects urine-related information and analyzes the information in real-time to make early predictions of urine infections. The key contributions of the proposed framework are listed as follows:

1. Introducing a novel solution for the prediction of urine infection using ensemble learning in the fog computing environment.
2. Fog computing-based Infectious Risk Factor (IRF) computation to determine the probability of the related disease in a given time window.
3. Deploying cloud layer for storage of user health records with the predicted results of urine infection for the longer time span.
4. Experimental analysis of the proposed framework to determine its effectiveness in a simulated environment.
5. Comparison of the presented framework with previously proposed methodologies to prove its effectiveness and novelty aspects.

1.3 Article Outline

The rest of the article is structured into different sections. Section 2 provides a review of recent literature concerning the current space of research. In Section 3, a modular approach for the early prediction of urine infection is discussed. Section 4 provides detail about the dataset and various performance evaluation metrics. Section 5 provides the experimental implementation and performance measures of the proposed framework. Finally, Section 6 concludes the article with some important directions for futuristic research.

2 Related Work

In this section, the extensive study of recent literature on novel technologies such as the Internet of Things and machine learning has been carried out to understand their use in the fog-cloud environment. Moreover, various research gaps have been identified to propose a new framework in the field of healthcare.

2.1 IoT Enabled Health Monitoring Systems

With the recent advancements in the research on embedded systems and mobile computing, the rapid proliferation of IoT-based medical devices is going on, and the integration of these medical devices supported by

Table 1: Urine Test Parameters [9]

Parameter	Risk range	Suspected disease
Glucose	$> 130\ mg/dL$	Diabetes, liver disease, brain tumor
Protein	$> 20\ md/dL$	Diabetes, kidney infections
Ketones	$> 40\ mg/dL$	Diabetes
Bilirubin	$> 1\ md/dL$	Hepatitis, liver disease
pH value	$< 4\ or > 9$	Gout, dehydration, UTI, diabetes
Specific gravity	> 1	Cystitis
Hemoglobin	$> 3\ RBCs$	Kidney tumor or inflammation
Leukocytes	$> 5\ wbc/hpf$	UTI, kidney disease
Nitrite	$>= 0$	Cystitis

mobile computing and Information and Communication Technologies (ICT) have developed numerous health monitoring systems. Moreover, the massive amounts of data generated by medical devices, and wireless sensors based on the Internet of Things provide huge opportunities for data analysis and remote patient monitoring in the healthcare field. In 2021, Albahri et al. [10] presented a state-of-the-art multi-field systematic review on IoT-based telemedicine for the promotion of health and disease prevention. The authors also highlighted several challenges, motivations and research directions for using the Internet of Things in a telemedicine environment. In 2021, Rustagi et al. [11] proposed a framework for emergency medical insurance services based on the Internet of Things. The framework aimed at communicating patient's critical medical details to the nearby hospitals and by creating a green corridor for ambulance crossings. In 2021, Hajjaji et al. [12] presented a systematic review on the big-data and Internet of Things-enabled technologies for smart environments. Several key challenges, applications, trends, and current research in integrating big data into smart environments (such as healthcare and other related disciplines) were also highlighted. In 2021, Ali et al. [13] proposed a healthcare framework based on the Internet of Things for early detection of dysphonia. The framework used the edge computing paradigm and high-order directional derivatives to analyze the time-frequency spectrum of the signal. The system was able to respond in minimal time and with a higher level of accuracy. In 2020, Hosseinzadeh et al. [14] proposed a health monitoring system for elderly patients based on the Internet of Things. The system was capable of remotely monitoring the patient's health by sensing their biologic and behavioral parameters through IoT devices embedded in the patient's proximity. In 2020, Otoom et al. [15] proposed a framework based on the Internet of Things for early detection and prediction of coronavirus cases. Several decision-making algorithms namely, Support Vector Machine (SVM), Neural Networks, Naive Bayes, K-Nearest Neighbor (K-NN), Decision Tree, OneR, and ZeroR were used to identify potential coronavirus cases. In 2019, Sengupta [16] emphasized the safety of patient monitoring systems and proposed an authentication and security system based on the Internet of Things. By using timestamps and biometric hash functions for previously available security mechanisms, the system had great potential to provide security against cyber attacks. In 2018, Verma and Sood [17] proposed an IoT-enabled framework for the remote monitoring of patients in smart homes. The proposed system used data mining technology and cloud storage and alert generation components at the edge of the network for efficient and real-time processing.

2.2 Machine Learning in IoT enabled Fog-Cloud Environment

In 2021, Kallel et al. [18] presented a COVID-19 prediction framework by integrating the Internet of Things in fog-cloud healthcare environments. The hybrid framework used a series of machine learning services, among which Stream-MLaaS was implemented in the fog layer for short-term decision-making, and the long-term decision was executed in the cloud layer by using Batch-MLaaS. In 2021, Kaur and Verma [19] conducted a comprehensive review of the Internet of Things in the fog-cloud health environment. The authors also highlighted several case studies such as Hypertension detection and future trends and design issues for the implementation of these technologies in healthcare. In 2021, Abdulkareem et al. [20] realized the use of machine learning in the fog computing environment to achieve the effectiveness of COVID-19 diagnosis and proposed an intelligent system. The model aimed at reducing the workload of medical staff and the mortality of patients during the coronavirus pandemic. In 2021, Rathor et al. [21] presented a survey on emotional health prediction based on machine learning and the Internet of Things. Facial expressions were recognized to provide information about

a user's emotional health. In 2021, Dejia and Hanzhong [22] proposed a model based on artificial intelligence and recurrent neural networks to predict the mortality of COVID-19 patients admitted to the ICU. In 2020, Kemal [23] incorporated a Growing and Pruning-based Deep Neural Network (GP-DNN) for the efficient diagnosis of Parkinson's disease. The simulation results showed that, compared with other models, the GP-DNN-based model has higher predictive performance. In 2020, a system for detecting arrhythmia was developed in cooperation with the Internet of Things using data mining technology in a fog environment by Moghadas et al. [24]. The authors used the k-nearest neighbor approach to classify and validate the model. In 2020, Banerjee [25] proposed a model that uses machine learning to make decisions in the intensive care unit (ICU) under the fog environment of the Internet of Things. The proposed model performed real-time processing by bringing the computation closer to the data source. Blockchain technology was also introduced to improve the security of the model. In 2020, Xing et al. [26] recommended the use of deep learning algorithms for the classification of COVID-19 cases. The authors also conducted various simulations, and the results show that the Keras deep learning classification algorithm outperforms SVM and CNN on various statistical indicators. In 2019, Verma and Sood [27] proposed a framework for the management of student stress in the fog-cloud Internet of Things environment. The Bayesian Belief Network (BBN) classifier and TBDN prediction models were used to evaluate the accuracy of the proposed framework.

2.3 Research Gaps

Recent research on the Internet of Things, fog-cloud computing, and machine learning has proposed a variety of novel solutions in the field of healthcare. Although a variety of health monitoring systems have been developed, we merely find any smart health framework that addresses the detection of urine infections in the home environment, particularly to the best of our knowledge. Considering these facts, we presented a framework for the early prediction of urine infections in smart homes by using the Internet of Things in a fog environment.

3 System Model

The layered architecture of the proposed framework for the early prediction of urine infections in the home environment has been presented in Figure 1. The proposed framework is composed of three layers namely, the data acquisition layer, fog layer, and cloud layer. Each layer of the proposed framework is entrusted with pre-defined roles and responsibilities that are necessary to achieve the objectives of the system. Multiple wireless sensors based on IoT are deployed in the data acquisition layer to collect the user's urine-related data. The incorporation of the fog computing paradigm enables real-time processing and efficient decision-making at the edge of the IoT devices, without causing any unnecessary delay. Moreover, the capabilities of cloud computing allow storage of predicted results along with patient's health records for a long duration of time. The task performed by each layer of the system model is provided in-depth in the subsequent sections.

3.1 Data Acquisition Layer

To efficiently predict and monitor urine infections in the home environment, accurate information about various urine parameters is required. Hence, a layer namely the data acquisition layer is introduced in the proposed architecture for data collection. The data acquisition layer is the first layer in the proposed framework that works closer to the user and is responsible for collecting the user's urine-related information such as blood-cell count, pH value, bilirubin, ketones, pus cells, etc. All this information is collected with the help of sensor-equipped small containers embedded in smart toilets. Various sensors related to urinalysis are incorporated with the smart toilets and are used for sensing and computation of different urine parameters such as color, odor, etc. without any human obstruction. For the identification of each user, different Radio frequency identification (RFID) tags or fingerprint sensors are attached to the flush system. Table 2 provides the detail of different types of sensors for urine analysis.

Table 2: Different Sensors for Urine Analysis

Sensor Type	**Purpose**
Electronic Hydrometer	Specific gravity
Bilirubin-optical Sensor	Bilirubin
Raspberry-Pi	Temperature, color
Catheter	pH value
Dipsticks	Nitrite

The information obtained using various sensors embedded in the smart toilet system is transmitted to the fog layer through the gateway device using advanced communication technologies (such as Wi-Fi, ZigBee,

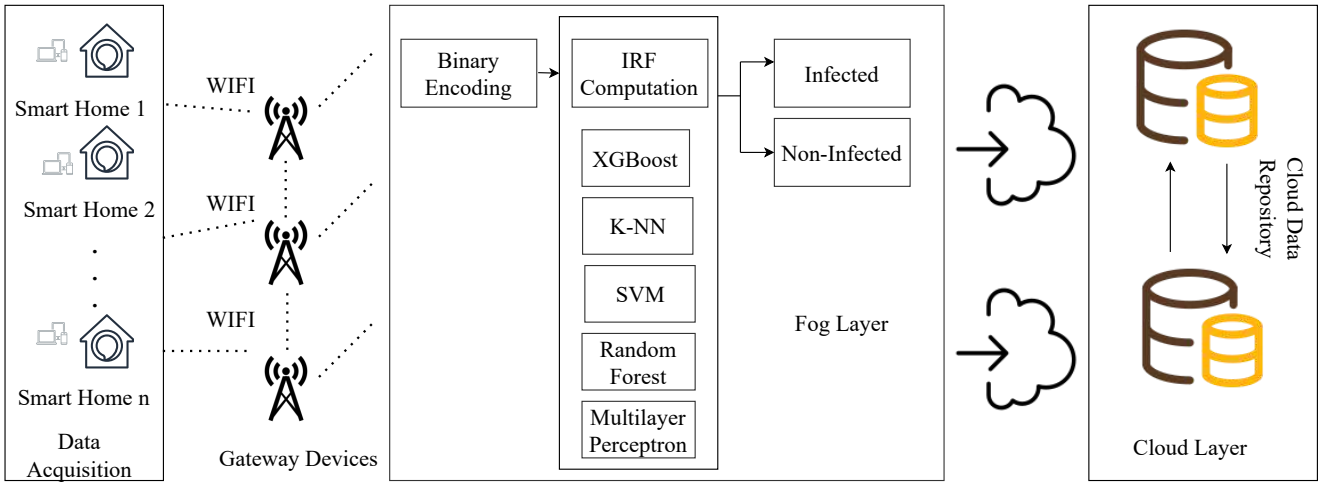


Fig. 1: Layered Architecture for Urine Infection Prediction

6LoWPAN, etc.) [28]. The fog layer uses these data for various data analysis and decision-making operations. Several protocols namely secure socket layer (SSL), and transmission control layer (TCL) may also be implemented for ensuring the security of data transmission to the forthcoming layers.

3.2 Fog Layer

The fog layer is responsible for analyzing urine-related data at the edge of the IoT device and providing decision-making by calculating infectious risk factors (IRF). The value of IRF is obtained using machine learning algorithms, and its value determines whether the patient is infected or uninfected. Fog layer consists of data preprocessing, followed by classification component. Each component in fog layer is explained ahead.

3.2.1 Data Preprocessing

Data preprocessing is one of the most critical steps before data analysis and decision-making. Therefore, before implementing different machine learning algorithms, the urine data obtained using IoT sensors must be preprocessed and converted into a form that is suitable for machine learning algorithms. In our proposed framework, a feature representation technique has been used for binary encoding of different urine-related parameters obtained through data acquisition layer. Values given in Table 1 is used to find the suitable encoding such that a zero value is assigned, if the respective parameter possesses a safe value. However, a 1 value is assigned if the value of the respective parameter crosses the safe threshold. For instance, in the preprocessed dataset, pH is assigned a value 1, if it lies within the range of < 4 or > 9, specific gravity has value 1, if it has value > 1 and similar encoding is performed for other parameters present in the dataset.

3.2.2 Data Classification

The preprocessed data is classified using different machine learning algorithms to check the presence of urine infection in humans by the analysis of different above-mentioned parameters. The fog layer is helpful in the timely analysis of the preprocessed data due to the proximity of the fog server near the home environment. The fog layer is also responsible for infectious risk factors (IRF) computation based on the classification of processed data. IRF is used to predict the probability of urine infection-related disease in a given time window. The level of IRF is used to find whether the patient is healthy or there are possibilities of any risk. We model our system in such a way that urine data is collected three times a day that includes morning, afternoon, and evening after every meal. This will help to analyze the patient's health continuously and hence reduces any type of risk by early detection of any anomalies.

Depending upon the probability prediction of IRF, we can divide the collected data into two classes namely, infectious class and non-infectious.

Infectious Class The data value of urine information where different parameters have crossed the safe zone limits come under the category of this class. Those users, for which data belongs to this category, must require immediate attention from medical staff.

Non-infectious Class Data values for which different parameters such as blood cell count, ketone, glucose,

etc. have values in the safe zone, comes under this category. If any user's urine data belong to this category, that means there is the absence of any type of infection risk.

The present study is designed in such a way that the following classification models are utilized to classify the urine-related data into different predefined classes:

(a) Extreme Gradient Boosting (XGBoost): In 2016, Chen et al. [29] introduced the extreme gradient boosting (XGBoost) algorithm. XGBoost is a decision tree-based ensemble learning technique that works for both regression and classification problems. Variety of applications of the XGBoost have been cited in many application domains and the approach works well for small-to-medium sized structured data [30]. Boosting is a machine learning method that sequentially adds several different weak learners to realize a strong learner with higher computing power [31]. Weak learners are added repetitively by using a gradient descent algorithm until a strong learner with noticeable performance is achieved. The gradient descent algorithm results in minimizing the loss and hence improves the performance of the predictive model. XGBoost can be realized as an extension to the gradient boosting framework by providing many features including tree-pruning using the depth-first approach, parallelization of weak learners, handling missing values, and a feature of regularization to overcome overfitting problems [32]. Several other interesting features that differ XGBoost from other machine learning algorithms are language independence, out-of-core computing, cache awareness, handling sparse data, weighted quantile sketch, etc.

The basic idea behind XGBoost is as follows:
Given a dataset $D_S = \{(x_i, y_i)\}$, where x_i specifies the range of urine-related attributes and y_i represents the predictive output as infected or non-infected class. A series of ensemble decision trees are generated in various iterations, where the last predicted residual in each iteration is considered to generate the objective function. In order to start a new iteration, a new fitting model is calculated by considering the residual fitting based on the first and second derivatives of the loss function matrix. This means that if we run p iterations, then we will generate p decision trees. The greedy method is used to find the best split point for the next iteration. At last, we have multiple trees generated, where each tree has several leaf nodes with the score of the sample. The final predicted value is calculated by multiplying the scores of various leaf nodes with their corresponding weights. In the XGBoost ensemble model, suppose we are iterating the model p times, i.e. at the end we will have p different trees then the final predicted value is computed as given in Equation 1.

$$\hat{y}_i = \sum_{p=1}^{p} f_p\left(x_i\right), f_p \in F \tag{1}$$

where, $\hat{y}_i$ is the final predicted value, $f_p(x_i)$ is the predicted value at p^{th} iteration using the residual tree, F is the function that represents the space of the residual tree. The objective function that needs to be minimized in each round is evaluated by using Equation 2 and has two main components-loss function and regularization.

$$f_p = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \Omega\left(f_p\right) \tag{2}$$

where, l is the loss function that computes the distance between y_i and $\hat{y}_i$
Ω represents the regularization term that results in avoiding the model overfitting.

The XGBoost model for prediction of urine infections can be optimized to achieve better performance. Therefore, a 10-fold cross validation technique has been utilized for training purpose. By using the 10-fold cross-validation technique, the model is trained on all available data. Moreover, this also results in minimizing the bias and hence avoids the overfitting problem. This technique splits the whole dataset into 10 folds and each fold comprise equal number of instances. Each fold serves as a test set to judge the performance of the model while the remaining folds are used for training purpose. This step is repeated until the model is trained and tested over all the folds. Moreover, to obtain the best hyperparameters for training the model, a tuning approach namely grid search has been employed. A grid search technique test the model with all possible combinations of different parameter values (learning-rate, max-depth, number of estimators, gamma value, reg-alpha, etc.) and selects the one which result in an optimal performance. The workflow of proposed XGBoost ensemble classifier is depicted in Figure 2.

(b) Support Vector Machine: In order to perform binary classification, a non-linear SVM model for data prediction has been utilized. SVM uses the concept of hyperplane to classify N-dimensional space. The difference is that the nonlinear version uses kernel functions such as RBF (radial basis function) and sigmoid. This type of model is very helpful to classify high-dimensional non-linear data [33].
The following Equation 3 and Equation 4 represents the equation of hyperplane and the distance between a point x and a hyperplane (β, β_0) respectively.

$$f(x) = \beta_0 + \beta^T x \tag{3}$$

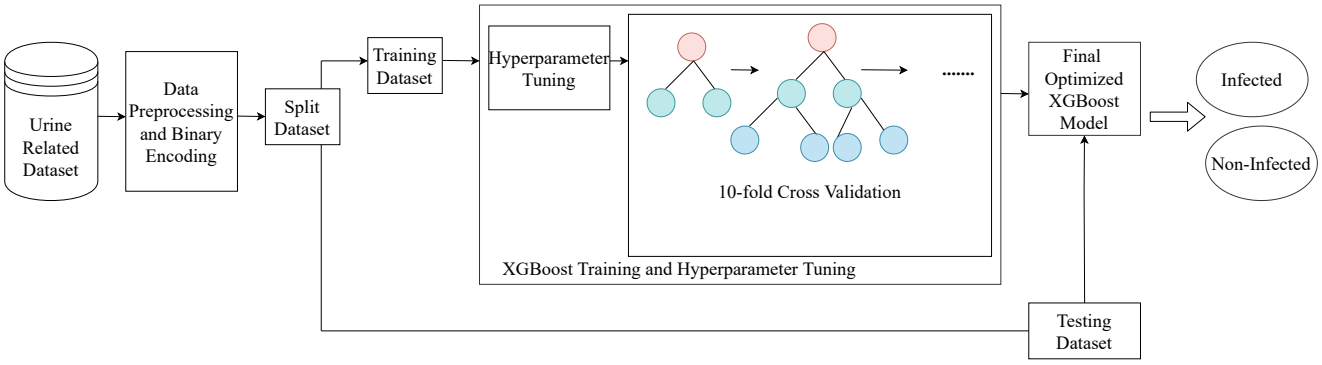


Fig. 2: Workflow of XGBoost Ensemble Classifier

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} . \quad (4)$$

where, β represents the weight vector.
β_0 specifies the bias.

(c) Random Forest: Random forest is regarded as a collection of many decision trees. RF is the classic example of a decision tree-based ensemble learning algorithm. Analogous to XGBoost, random forest can also be used for classification and regression, but the ground technology used in random forest is bagging, while XGBoost uses boosting. The main disadvantage of random forest is the slow prediction speed due to the generation of a large number of sequential decision trees, making it unsuitable for any practical applications [34].

(d) K-Nearest Neighbor: In the k-nearest neighbor algorithm, any sample is classified by calculating the distance to the nearest neighbor. This is one of the simplest and easy-to-implement classification techniques. To find the best value of k, we start our algorithm with k=1 and increase it gradually. From the results, it is evident that the optimal results are obtained when the results are computed by using the value of k=3.

(e) Multilayer Perceptron: The Multi-layer Perceptron (MLP) is one of the most important classes of feed-forward artificial neural networks. One of the several features of MLP includes classification. MLP is a versatile algorithm that has been implemented in a variety of healthcare applications [35]. A Multi-layer perceptron model comprises an input layer, an output layer, and one or more hidden layers. Each layer in MLP consists of a collection of artificial neurons. Input layer takes as input a set of urinary infections relevant attributes while the output layers are responsible for the final decision on one of the two classes infectious or non-infectious. The collection of one or more hidden layers forms the computational part of MLP. In the current research, an MLP has been implemented using 12 input nodes that correspond to the input features. The output layer consists of 2 nodes for classifying the result as either infected or non-infected. Moreover, several other parameters such as learning rate, number of hidden layers have been adjusted to provide better results.

3.2.3 Cloud Layer

This is the last layer of the proposed framework for the early prediction of urine infection in smart home environment. The main purpose of this layer is to store all historical results of urine infection prediction in the data repositories present in the cloud. Although, numerous cloud service providers are available for performing computation, however, the far locality of cloud data centers may result in a delayed response with increased bandwidth. Therefore, in our proposed framework, a fog environment has been employed for the prediction and analysis and the cloud layer only works as the repository to store the predicted results along with patient's health records.

4 Dataset Description and Performance Measures

4.1 Data Acquisition

In order to test the performance of the proposed framework, a dataset related to urinary infections is needed. To address this issue, a rigorous search across multiple data repositories (UCI, Kaggle, Dataworld, etc.) was performed. However, no dataset specifically for urine infection was found on the Internet. Therefore, the latest data of nearly 50 patients (32 females and 18 males) is retrieved from nearby pathology laboratories. To validate the performance of the system, the dataset with a limited number of instances is not sufficient. Hence,

a bootstrap technique [36] is applied for generating instances. The bootstrap method is a resampling technique that involves iteratively generating new samples with a replacement policy. It is a popular machine learning technique for dealing with smaller datasets and has been cited in many papers [37][38]. The data of 50 patients is bootstrapped to 8000 patients with equal probability of consideration of all the possible samples. To maintain the patient's privacy, the name attribute is replaced with a unique identification number (UID). The resampled dataset is then binary encoded to convert all parameter values to 0 and 1. The dataset so obtained has been used for classification purposes by employing different machine learning algorithms. A sample urine dataset with all the possible combinations is presented in Table 3 where gl, pr, kt, bl, pH, sg, lk, and nt represents glucose, protein, ketone, bilirubin, pH value, specific gravity, leukocytes, and nitrite respectively.

4.2 Performance Metrics

The performance of the proposed framework has been evaluated against various parameters as:

1. Accuracy: Accuracy is a measure of the number of correctly classified instances to the total number of instances in the dataset. The measure of accuracy can be mathematically expressed as:

$$\text{Accuracy } = \frac{TP + TN}{TN + TP + FP + FN} \tag{5}$$

2. Sensitivity: It is defined as the measure of the ability of a proposed model to correctly detect the patients with urine infections and mathematically expressed as:

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{6}$$

3. Specificity: It is defined as the measure of the ability of a proposed model to correct identify the patients with no possible cases of urine infections and is mathematically expressed as:

$$\text{Specificity } = \frac{TN}{TN + FP} \tag{7}$$

4. Precision: It defines the proportion of data points that are classified as infected and are actually infected.

$$\text{Precision } = \frac{TP}{TP + FP} \tag{8}$$

5. F-Score is obtained by using the results of precision and recall obtained in the data analytics and mathematical formula for the calculation of F-score is given in equation below.

$$F - Score = \frac{2 \times \text{ precision } \times \text{ recall}}{\text{precision } + \text{ recall}} \tag{9}$$

6. Classification Time: It is the measure of the time taken by the classifier to classify the result as infected or non-infected.

Where, the values of TP, TN, FP, FN denotes the true positive, true negative, false positive, and false negative rates respectively. True positives are data samples correctly classified as infected; false positives are those data samples that are not-infected but predicted as infected; true negatives are data samples correctly predicted as non-infected and finally, false negatives represent data samples which are infected but classified as non-infected. More specifically, the relationship between the actual class and the predicted class is depicted in Figure 3.

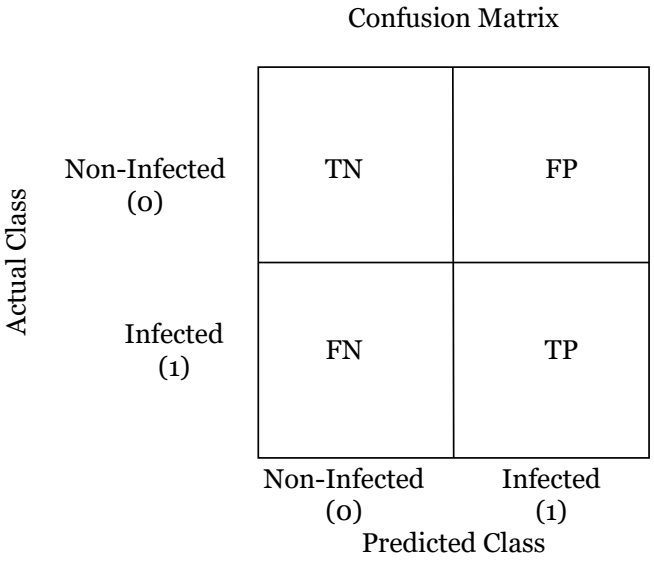


Fig. 3: Confusion Matrix for Binary Classification

5 Experimental Setup

In this section, the proposed framework has been implemented in a simulated environment and various performance metrics (accuracy, specificity, sensitivity, precision, and f-score) have been calculated to evaluate the efficacy of the proposed framework. The experimentation is conducted on a computer with following specifications: Intel(R) Core(TM) i3 processor, a primary

Table 3: Dataset for Urine Test Samples

UID	Age	Gender	Color	Appearance	gl	pr	kt	bl	pH	sg	lk	nt
1	56	Male	colorless	hazy	0	0	1	1	0	0	0	1
2	42	Female	yellow	clear	0	0	1	1	1	0	1	1
3	35	Female	red	cloudy	1	1	0	1	1	0	1	1
4	27	Male	colorless	clear	0	0	0	1	0	1	0	0
5	40	Male	amber	hazy	1	1	0	1	0	0	1	0
6	65	Female	light yellow	hazy	0	1	0	1	0	1	1	0
7	48	Female	yellow	cloudy	0	0	1	1	1	0	1	1
8	40	Female	colorless	clear	0	1	0	1	1	0	1	1
9	29	Female	dark yellow	cloudy	0	1	0	1	1	0	1	1
10	22	Male	brown	hazy	1	0	1	1	1	0	1	1

memory capacity of 4 GB, clock frequency of 2.10 GHz and 64 bit Windows-8 Operating system.

5.1 Training of XGBoost Ensemble Model

The proposed framework uses a XGBoost classifier based on ensemble learning technique for binary classification of users into infected or non-infected class. The model is trained using xgboost module available in the latest version of python. In order to achieve optimized performance, the values of different parameters are tuned using a hyperparameter tuning technique and then a 10-fold cross-validation technique has been utilized for training the model. Algorithm 1 describes the working of 10-fold cross-validation technique. The proposed system uses a grid search-based optimization technique [39] to select the best possible combination of different XGBoost parameter values. Various parameters considered for parameter tuning are presented in Table 4. After successful training of the XGBoost model, it is tested against a set of performance metrics to evaluate its predictive efficiency.

Algorithm 1: Working of 10-Fold Cross Validation

1: Randomize the dataset.
2: Split the dataset into 10 equal folds.
3: **for** each fold **do**
4: Train the model by combining 9 folds, taking out one fold as a test fold.
5: Test the performance of the model using the test fold.
6: Perform the tuning of the classifier parameters.
7: Compute the statistical scores.
8: **end for**
9: Return the model's performance by averaging the statistical scores of different folds.
10: Exit

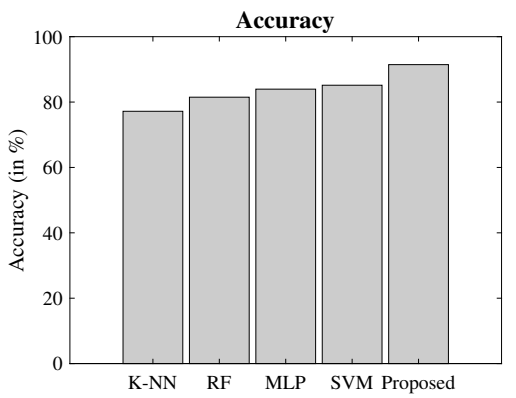
Table 4: XGBoost Optimal Parameter Values

Hyperparameter	Value
Learning rate	0.10
Max tree depth	4
Estimators	100
Gamma value	7
Alpha value	0.2
Objective	binary:logistic
Evaluation metric	Error

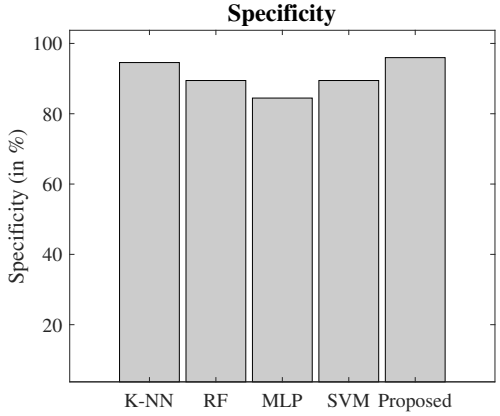
5.2 Results and Discussions

In this section, the results obtained in the simulation process are discussed to verify the effectiveness of the proposed strategy. The test results of the proposed methodology are depicted in Figure 4, where, Figure 4 (a) depicts the graphical representation of the accuracy of the proposed model in comparison to other models. The results of specificity, sensitivity, precision, and f-score are plotted in Figure 4 (b), Figure 4 (c), Figure 4 (d), and Figure 4 (e) respectively. Apart from the statistical measures, classification time results of different algorithms are also computed and are shown in Figure 4 (f). It is evident from the results that performance of the proposed XGBoost model seems better from all evaluation aspects.

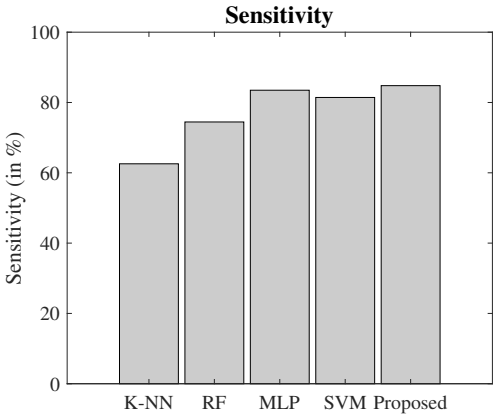
Moreover, two comparative studies have been conducted to evaluate the effectiveness of the proposed framework. First, the statistical results for different machine learning algorithms are computed by using the same dataset and computed results are compared with the proposed XGBoost ensemble model. Furthermore, the proposed framework is compared with previously proposed methods in the current research problem to prove its utility for real-time scenarios.



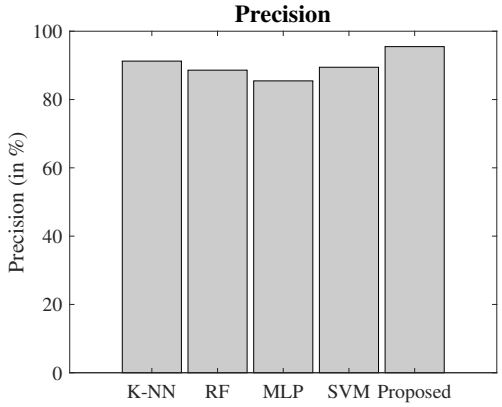
(a) Accuracy



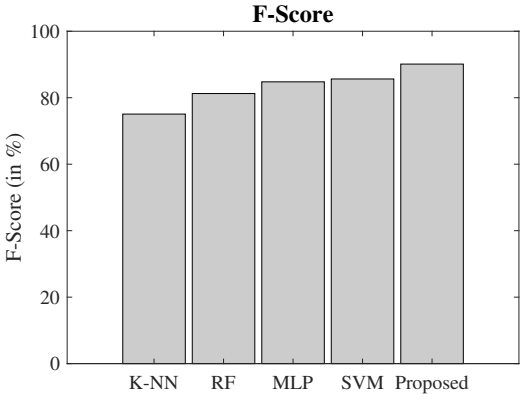
(b) Specificity



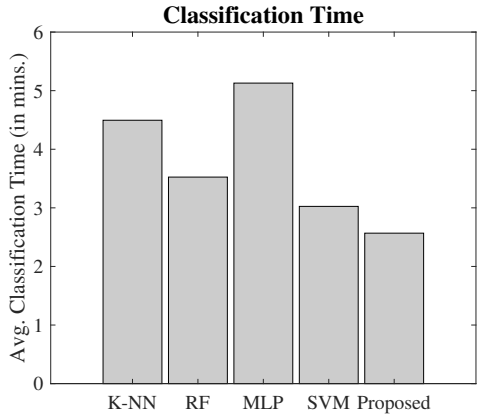
(c) Sensitivity



(d) Precision



(e) F-Score



(f) Classification Time

Fig. 4: Comparative analysis of performance based on Urinary Infections dataset
(a) Accuracy; (b) Specificity; (c) Sensitivity; (d) Precision; (e) F-Score; (f) Classification time

5.2.1 A Comparative Study with other Machine Learning Algorithms

To evaluate the effectiveness of any framework, it is necessary to evaluate its performance based on a set of baseline algorithms. Various baseline algorithms considered for comparative analysis in the current research include k- nearest neighbor (K-NN), random forest (RF), multi-layered perceptron (MLP), and support vector machine (SVM). The testing results obtained by the utilization of different machine learning techniques are reported in Table 5. It is evident from the results that

the proposed XGBoost ensemble learning technique outperforms all other algorithms by achieving a classification accuracy of 91.45 %. Moreover, the test results of specificity, sensitivity, precision, and F-score also validate the performance improvement of the XGBoost ensemble model over other classification techniques.

5.2.2 A Comparative Study with other Recent Scholarly Works

In this section, we compare our proposed work with recently published work on predicting urine infections. Several factors considered for carrying out the comparative analysis include the Internet of Things (IoT), Fog computing (Fog), Cloud computing (Cloud), Home-centric environment (Home), Decision-making approach (Method), and finally, predictive-accuracy (Accuracy). Table 6 presents the comparative analysis of the previous works with our proposed methodology. In [40], the authors attained the highest accuracy of 98.3 %, however, the experimentation was performed with very few attributes. Moreover, the presented technique did not incorporate an IoT-fog-cloud (IFC) architecture to process the results in real time. In [41], the authors provided a model based on the XGBoost technique for the prediction of urine infections. However, the number of patients considered for the study was very small. Moreover, the home environment was also not considered as a part of the predictive framework which limits the utility of the proposed approach. In [42], the authors presented a case study of patients with a urine infection. The authors considered only the elder patients to conduct the study. However, the performance estimator of accuracy was not addressed by the authors in the proposed study. Based on this comparative analysis, it can be concluded that the strategy presented in the current study shows notable improvement and performs significantly better than the previously proposed methods in all aspects.

6 Conclusions and future scope

Urine infection is one of the most prevalent health concerns in modern society. With the rapidly changing lifestyle, the cases of urine infections are frequently increasing. In this paper, a generic framework for the early diagnosis and prediction of urine infections has been presented by utilizing the capabilities of the Internet of Things in a fog environment. In the proposed framework, wireless sensors based on the Internet of Things have been used for data acquisition, followed by data analysis at the fog layer. Initially, data obtained using multiple sensors is preprocessed and Infectious Risk Factor (IRF) is calculated using the XGBoost ensemble learning approach at the fog layer. The IRF determines the probability of urine infection-related disease in a given time window. Depending upon the IRF value, a patient is classified into one of the two classes-Infectious and Non-infectious. Correspondingly, the results of the prediction along with the patient's details are stored in the cloud repository for future analysis. The proposed framework based on the XGBoost ensembling technique has been extensively tested using real-time patient data. Statistical results of accuracy, sensitivity, specificity, precision, and f-score averaging to 91.45 %, 84.79 %, 95.96 %, 95.49 %, and 90.12 % respectively prove the effectiveness of the proposed methodology for early prediction of urine infections. Moreover, the classification time of the proposed methodology also justifies its implementation for real-world scenarios.
In the future, multiple ensemble techniques available in the literature can also be implemented to check the reliability of the proposed methodology. Moreover, the security of collected data and efficient storage of predicted results in the cloud against unauthorized access is the another concern for further exploration in future research.

Conflict of Interests

On behalf of all the authors, corresponding author declares that there is no conflict of interest involved in conducting the study.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Chunguang Bai, Patrick Dallasega, Guido Orzes, and Joseph Sarkis. Industry 4.0 technologies assessment: A sustainability perspective. *International journal of production economics*, 229:107776, 2020.
2. Mohd Javaid, Abid Haleem, Raju Vaishya, Shashi Bahl, Rajiv Suman, and Abhishek Vaish. Industry 4.0 technologies and their applications in fighting covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):419–422, 2020.
3. Rajeev Kumar, Vibha Jain, Naveen Chauhan, and Narottam Chand. An adaptive prediction strategy with clustering in wireless sensor network. *International Journal of Wireless Information Networks*, 27(4):575–587, 2020.

Table 5: Comparative Performance among different Predictive Models

Predictive Model	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**F-score**
XGBoost	0.9145	0.9596	0.8479	0.9549	0.9012
Random Forest	0.8149	0.8945	0.7446	0.886	0.8126
MLP	0.8395	0.8446	0.8349	0.8546	0.8478
K-NN	0.7718	0.9456	0.6256	0.9125	0.7508
SVM	0.8515	0.8945	0.8143	0.8945	0.8564

Table 6: A Comparative Study with other Recent Works

Study (Year)	**IoT**	**Fog**	**Cloud**	**Home**	**Method**	**Accuracy**
Ilker et al. [40] (2018)	NO	NO	NO	NO	ANN	98.3 %
Taylor et al. [41] (2018)	NO	NO	NO	NO	XGBoost	87.5 %
Marilyn et al. [42] (2011)	YES	NO	NO	NO	Fuzzy	**
Proposed	YES	YES	YES	YES	XGBoost	91.45%

4. Syed Wasif Abbas Hamdani, Abdul Waheed Khan, Naima Iltaf, Javed Iqbal Bangash, Yawar Abbas Bangash, and Asfandyar Khan. Dynamic distributed trust management scheme for the internet of things. *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(2):796–815, 2021.
5. Vibha Jain and Bijendra Kumar. Combinatorial auction based multi-task resource allocation in fog environment using blockchain and smart contracts. *Peer-to-Peer Networking and Applications*, pages 1–19, 2021.
6. G Athinarayanan, R Mariselvam, P Dhasarathan, AJA Ranjitsingh, et al. Epidemiology of urinary tract infection in south india. *World Journal of Biology Pharmacy and Health Sciences*, 1(1):025–032, 2020.
7. Martha Medina and Edgardo Castillo-Pino. An introduction to the epidemiology and burden of urinary tract infections. *Therapeutic advances in urology*, 11:1756287219832172, 2019.
8. Katie Rutledge-Taylor, Anne Matlow, Denise Gravel, Joanne Embree, Nicole Le Saux, Lynn Johnston, Kathryn Suh, John Embil, Elizabeth Henderson, Michael John, et al. A point prevalence survey of health care-associated infections in canadian pediatric inpatients. *American journal of infection control*, 40(6):491–496, 2012.
9. Joris Penders, Tom Fiers, and Joris R Delanghe. Quantitative evaluation of urinalysis test strips. *Clinical chemistry*, 48(12):2236–2241, 2002.
10. AS Albahri, Jwan K Alwan, Zahraa K Taha, Sura F Ismail, Rula A Hamid, AA Zaidan, OS Albahri, BB Zaidan, AH Alamoodi, and MA Alsalem. Iot-based telemedicine for disease prevention and health promotion: State-of-the-art. *Journal of Network and Computer Applications*, 173:102873, 2021.
11. Aanshi Rustagi, Mansi Shukla, FCD Samuel, S Ananda Kumar, Amit Banerjee, Sangeetha Ramaswamy, and L Ramanathan. Data analysis and interpretation in iot-based systems for critical medical services and healthcare applications. *Wireless Personal Communications*, pages 1–16, 2021.
12. Yosra Hajjaji, Wadii Boulila, Imed Riadh Farah, Imed Romdhani, and Amir Hussain. Big data and iot-based applications in smart environments: A systematic review. *Computer Science Review*, 39:100318, 2021.
13. Zulfiqar Ali, Muhammad Imran, and Muhammad Shoaib. An iot-based smart healthcare system to detect dysphonia. *Neural Computing and Applications*, pages 1–11, 2021.
14. Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Marwan Yassin Ghafour, Aram Mahmood Ahmed, Parvaneh Asghari, Alireza Souri, Hamid Pourasghari, and Aziz Rezapour. An elderly health monitoring system based on biological and behavioral indicators in internet of things. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–11, 2020.
15. Mwaffaq Otoom, Nesreen Otoum, Mohammad A Alzubaidi, Yousef Etoom, and Rudaina Banihani. An iot-based framework for early identification and monitoring of covid-19 cases. *Biomedical Signal Processing and Control*, 62:102149, 2020.
16. Sushanta Sengupta. A secured biometric-based authentication scheme in iot-based patient monitoring system. In *Emerging Technology in Modelling and Graphics*, pages 501–518. Springer, 2020.
17. Prabal Verma and Sandeep K Sood. Fog assisted-iot enabled patient health monitoring in smart homes. *IEEE Internet of Things Journal*, 5(3):1789–1796, 2018.
18. Ameni Kallel, Molka Rekik, and Mahdi Khemakhem. Hybrid-based framework for covid-19 prediction via federated machine learning models. 2021.
19. Karandeep Kaur and Harsh Kumar Verma. The interoperability of fog and iot in healthcare domain: Architecture, application, and challenges. In *Fog Computing for Healthcare 4.0 Environments*, pages 535–561. Springer, 2021.
20. Karrar Hameed Abdulkareem, Mazin Abed Mohammed, Ahmad Salim, Muhammad Arif, Oana Geman, Deepak Gupta, and Ashish Khanna. Realizing an effective covid-19 diagnosis system based on machine learning and iot in smart hospital environment. *IEEE Internet of Things Journal*, 2021.
21. Aartidevi S Rathor, Kirit Modi, and Makhduma Saiyad. A comprehensive survey on emotion based health prediction using internet of things and machine learning. In *Proceedings of the Second International Conference on Information Management and Machine Intelligence*, pages 173–182. Springer, 2021.
22. Hanzhong Zheng Dejia Shi. A mortality risk assessment approach on icu patients clinical medication events using deep learning. *Computer Modeling in Engineering & Sciences*, 128(1):161–181, 2021.

23. Kemal Akyol. Growing and pruning based deep neural networks modeling for effective parkinson's disease diagnosis. *Computer Modeling in Engineering & Sciences*, 122(2):619–632, 2020.
24. Ehsan Moghadas, Javad Rezazadeh, and Reza Farahbakhsh. An iot patient monitoring based on fog computing and data mining: Cardiac arrhythmia usecase. *Internet of Things*, 11:100251, 2020.
25. Anwesha Banerjee, Bhabendu Kumar Mohanta, Soumyashree S Panda, Debasish Jena, and Srichandan Sobhanayak. A secure iot-fog enabled smart decision making system using machine learning for intensive care unit. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–6. IEEE, 2020.
26. Xing Deng, Haijian Shao, Liang Shi, Xia Wang, and Tongling Xie. A classification–detection approach of covid-19 based on chest x-ray and ct by using keras pretrained deep learning models. *Computer Modeling in Engineering & Sciences*, 125(2):579–596, 2020.
27. Prabal Verma and Sandeep K Sood. A comprehensive framework for student stress monitoring in fog-cloud iot environment: m-health perspective. *Medical & biological engineering & computing*, 57(1):231–244, 2019.
28. Burak H Çorak, Feyza Y Okay, Metehan Güzel, Şahin Murt, and Suat Ozdemir. Comparative analysis of iot communication protocols. In *2018 International symposium on networks, computers and communications (ISNCC)*, pages 1–6. IEEE, 2018.
29. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
30. Duyen Thi Do and Nguyen Quoc Khanh Le. Using extreme gradient boosting to identify origin of replication in saccharomyces cerevisiae via hybrid features. *Genomics*, 112(3):2445–2451, 2020.
31. Talha Karadeniz, Gül Tokdemir, and Hadi Hakan Maraş. Ensemble methods for heart disease prediction. *New Generation Computing*, pages 1–13, 2021.
32. Nguyen Quoc Khanh Le, Duyen Thi Do, Fang-Ying Chiu, Edward Kien Yee Yapp, Hui-Yuan Yeh, and Cheng-Yu Chen. Xgboost improves classification of mgmt promoter methylation status in idh1 wildtype glioblastoma. *Journal of Personalized Medicine*, 10(3):128, 2020.
33. Syed Aziz Shah, Aifeng Ren, Dou Fan, Zhiya Zhang, Nan Zhao, Xiaodong Yang, Ming Luo, Weigang Wang, Fangming Hu, Masood Ur Rehman, et al. Internet of things for sensing: A case study in the healthcare system. *Applied sciences*, 8(4):508, 2018.
34. Pavleen Kaur, Ravinder Kumar, and Munish Kumar. A healthcare monitoring system using random forest and internet of things (iot). *Multimedia Tools and Applications*, 78(14):19905–19916, 2019.
35. Hongmei Yan, Yingtao Jiang, Jun Zheng, Chenglin Peng, and Qinghui Li. A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications*, 30(2):272–281, 2006.
36. Xuan Bao, Paramvir Bahl, Aman Kansal, David Chu, Romit Roy Choudhury, and Alec Wolman. Helping mobile apps bootstrap with fewer users. In *proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 491–500, 2012.
37. Prabal Verma, Sandeep K Sood, and Harkiran Kaur. A fog-cloud based cyber physical system for ulcerative colitis diagnosis and stage classification and management. *Microprocessors and Microsystems*, 72:102929, 2020.
38. Anil K Jain and JV Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987.
39. Ruiming Liu, Erqi Liu, Jie Yang, Ming Li, and Fanglin Wang. Optimizing the hyper-parameters for svm by combining evolution strategies with a grid search. In *Intelligent Control and Automation*, pages 712–721. Springer, 2006.
40. Ilker Ali Ozkan, Murat Koklu, and Ibrahim Unal Sert. Diagnosis of urinary tract infection based on artificial intelligence methods. *Computer methods and programs in biomedicine*, 166:51–59, 2018.
41. R Andrew Taylor, Christopher L Moore, Kei-Hoi Cheung, and Cynthia Brandt. Predicting urinary tract infections in the emergency department with machine learning. *PloS one*, 13(3):e0194085, 2018.
42. Marilyn J Rantz, Marjorie Skubic, Richelle J Koopman, Lorraine Phillips, Gregory L Alexander, Steven J Miller, and Rainer Dane Guevara. Using sensor networks to detect urinary tract infections in older adults. In *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services*, pages 142–149. IEEE, 2011.