IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 September 2022, accepted 27 September 2022, date of publication 3 October 2022, date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211264

RESEARCH ARTICLE

An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System

RADWA MARZOUK[1], ALA SALEH ALLUHAIDAN[2], AND SAHAR A. EL_RAHMAN[3], (Senior Member, IEEE)

[1]Department of Mathematics, Faculty of Science, Cairo University, Giza 12613, Egypt
[2]Department of Information Systems, College of Computer and Information Science, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[3]Electrical Engineering Department, Faculty of Engineering-Shoubra, Benha University, Cairo 13518, Egypt

Corresponding author: Ala Saleh Alluhaidan (asalluhaidan@pnu.edu.sa)

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R234), Princess Nourah bint Ab-dulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Diabetes, in all of its types, costs countries of all income levels unacceptably enormous personal, societal, and economic expenses. To predict type-2 diabetes, this work aimed to develop an analytical predictive model based on machine learning techniques and a web-based personalized diabetes monitoring system. The history of a patient will be collected and ready for analysis purposes based on machine learning techniques by continuously monitoring the patient's vital data. A diabetes monitoring system is proposed by utilizing the patient's QR card that allows the patients and doctors to be connected to the Internet of Things. So, they can deliver real-time information (such as insulin records) about their health status and can visit different healthcare institutions. The proposed system can help doctors to make data-driven decisions and enhance patients' treatment. Several machine learning algorithms that are Decision Tree, Support Vector Classifier, Random Forest, Gradient Boosting, Multi-layer Perceptron's, Artificial Neural Network, k-Nearest Neighbors, Logistic Regression, and Naive Bayes are used. The proposed analytical model is evaluated based on two different datasets a synthetic dataset and PIMA Diabetes Dataset. The performance of the classification models was analyzed in terms of accuracy, recall, and precision based on the cross-validation strategy. The findings show that the ANN model has better prediction accuracy than other models. The evaluation findings are analyzed and compared with other existing models. The system has dashboard graphs displaying the number of patients in Saudi Arabia cities. It also contains visualized graphs that include more detailed classifications for patients' states (Normal, Pre-Diabetes, and Diabetes).

INDEX TERMS Artificial neural network, data analytics, data visualization, diabetes, Internet of Things (IoT), machine learning, algorithms, QR code.

I. INTRODUCTION

Diabetes is a very common chronic disease in which blood sugar or blood glucose levels reach high. The three major types of diabetes are: Type 1, Type 2 and Gestational diabetes. Type 1 occurs when the human body cannot make enough insulin. Type 2 is a common type in about 90 percent of the patient where the human body cannot produce or use insulin properly. Gestational diabetes will happen during pregnancy and usually disappears after that [1]. Diabetes is the main cause of serious health problems such that eye disease, miscarriages, heart attack, kidney disease, and nervous system troubles.

There is no cure for Diabetes disease, but with the healthy diet and exercise a patient can control it. If diet and exercise are not sufficient, insulin treatment is needed. The major indication of Diabetes is frequent urination, increased thirst, and hunger. One of the vital medical problems is the detection

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Fadda.

of diabetes at early stage [2]. Early diagnosis of Diabetes diseases help to minimize the risk of patients having more complex health issues and medical costs. In the world of growing data, hospitals are deliberately adopting big data technologies. There are many advantages of utilizing data analytics in healthcare technologies to improve diagnosis, perfect outcomes, and minimize costs. Internet of things (IoT) technologies improve the healthcare system with high accuracy, lower cost, and save time. IoT is the Collection of objects connected to the internet which able to intercommunicate with each other. It can notify the hospital staff based on the patient's vitals. In addition, the patient's overall health can be viewed through prototypes of next-generation emergency care. Many researchers and engineers are interested in big data and IoT technologies to advance the next generation of smarter connected devices. Also, the data of embedded medical devices can be accessed anywhere in the world, and healthcare outcome is shared by hospitals, clinics, and health organizations [3], [4].

Machine learning techniques have an efficient accuracy rate in classification as compared with current classification algorithms. Machine learning aims to develop algorithms that make computer programs access and learn data by intelligence way. ML is the main branch of Artificial Intelligence (AI) and heavily relies on statistics. For more definitions of the ML see [5]. ML has solved several complicated and complex real-world problems in the application areas. Also, ML techniques have been used for the prediction and diagnosis of various diseases. It will be able to understand the input data, prediction, and make decisions. So, the researchers are developing many machine learning systems for diabetes prediction.

In this study, the proposed diabetes classification and monitoring system is developed to help the doctors to manage the time and make decision about the patient status. The system facilitates the secure data transferring and reducing the cost for patients based on the QR card. Also, patients can follow up their health condition continuously remotely. The efficiency of the proposed methodology is proved and the performance of the analytical proposed models can be used for prediction with satisfactory results and achieve improved rating optimization.

The paper is organized as follows: Section 2 presents a literature review. Section 3 covers the details of the machine learning algorithms. Section 4 describes the methodology. Section 5 shows the obtained results. Section 6 has the conclusion.

II. RELATED WORK

A literature review shows various results on diabetes diseases conducted by different methods and tools of diabetes diseases. Researchers have established various prediction models using ML to predict diabetes diseases. Salian and Harisekaran [6] suggest a system to identify the critical causes that cause readmission to diabetic patients. Predictive modeling has been used by employing the decision tree classification method. Authors in [7] analyzed diabetic treatment in the healthcare system utilizing big data analytics. Designing a predictive analysis system for diabetic treatment will improve data and give the greatest results in healthcare. For the prediction of diabetic diseases in pregnant women, the authors used Decision Tree and Naïve Bayes for prediction using the Hadoop / Map Reduce environment [8]. Sadhana et al. [9] analyzed the Pima Indians Diabetes Diseases dataset using a proposed structure that comprised Hive and R which is a phase of the big data concept.

The diabetic data warehouse was formed by an integrated healthcare system in the New Orleans area including 30,383 diabetic patients. Classification and regression tree approaches were used to analyze the above-mentioned dataset [10]. Saparkhojayev and Mukasheva give a short description of how Big Data technology is used in medicine, especially for diagnosing diabetes. Kopitar et al. used a machine learning prediction model for type 2 diabetes diseases. Early prediction and improvement clinical prediction [11]. Authors in [12] designed a system for diabetes disease classification via Support Vector Machine (SVM). They used Pima Indian diabetes dataset for proving the validatity of their approach. The accuracy was 78% based on the Radial Basis Function (RBF) kernel of SVM as the classifier model.

Krishnaveni and Sudha [13] proposed different techniques for predicting diabetic disease. They used discriminant analysis, KNN Algorithm, Naïve Bayes, SVM with linear kernel function, and SVM with RBF kernel function. The result proves the superior of Naïve Bayes. Naz and Ahuja created a multilayer feedforward neural network algorithm using deep learning for early prediction effectively. The algorithm attained a 98.07% accuracy rate in analyzing diabetes [14]. The authors introduced a diabetes risk prediction model using enhanced DNN. The model can detect whether someone will have diabetes disease in the future [15]. Alharbi and Alghahtani [16] established a hybrid model based on a genetic algorithm and Extreme LM algorithm to diagnose type 2 diabetes disease. The accuracy of that model was 97.5%. Harleen and Pankaj [17], [16] suggested a system based on data mining for predicting diabetes disease. The proposed system was evaluated with J48 and Naive Bayes. The achieved accuracies are 73.8% and 76.3%. Dimitrov [18] (see Figure 1) proposed mIoT as a critical piece of the digital healthcare transformation. The paper reviews mIoT and big data within healthcare. One of the best features of IoT in healthcare is the remote health monitoring system.

III. MACHINE LEARNING ALGORITHMS

The use of ML and its applications had grown exponentially in recent times. ML is a method of computer algorithms that improve automatically with more data and training.

A. RANDOM FOREST

The RF algorithm is a type of Classification and Regression methods that is formed via combining decision trees. Decision trees are easy to build, use, and interpret. RF combines

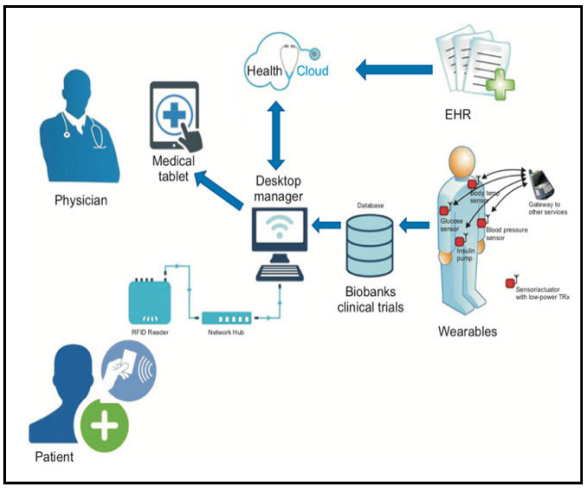


FIGURE 1. mIoT system [18].

the simplicity of decision trees with flexibility resulting in a huge improvement in accuracy. RF can handle large datasets. The trees were built using the classification methodology and the gradient trees. In tree group construction, RF uses two types of randomization: first, each tree is planted using a part of the training data. The second part of randomization is added when cultivating the tree by selecting a random sample of predictors in each node to select the best split [19].

The number of predictors specified in each node and the number of trees in the group are the two main parameters of the RF algorithm. RF developers have stated that the method does not require much synthesis of parameters and the default values usually generate good results for many problems. Once the forest is built, a new instance of a class is assigned by collecting trees, using a majority vote. Because of using a sample of boot training data, a third of the samples are deleted when each tree is constructed. These are called outside samples that can be used to evaluate workbook performance and build important measures [20]. A random forest is a meta estimator that fits a number of decision tree classifiers on many subsamples of the dataset and use averaging techniques to improve the prediction accuracy and control overfitting [20].

We can summarize RF algorithm as the following:

1- Chose random samples from a given dataset.
2- Build a decision tree for every sample. Then get the prediction result from every decision tree.
3- Vote for every predicted result.
4- Chose the greatest voted prediction result since the last prediction result. See Figure (2).

B. K-NEAREST NEIGHBORS

KNN is one of the simplest supervised ML algorithms that used to solve classification and regression methods. KNN supposes the similarity between the new data and available data and place the new data into the category that is most similar to the available categories. When new data appears

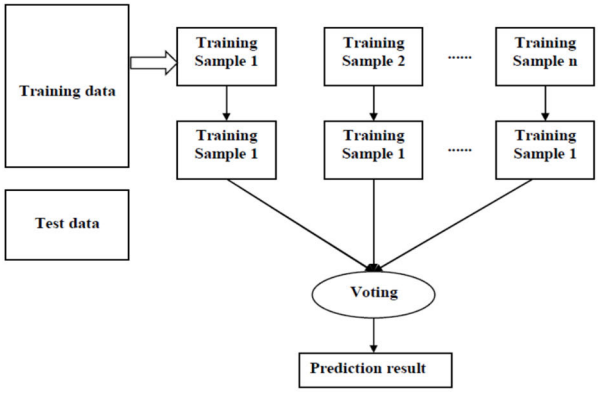


FIGURE 2. Random forest.

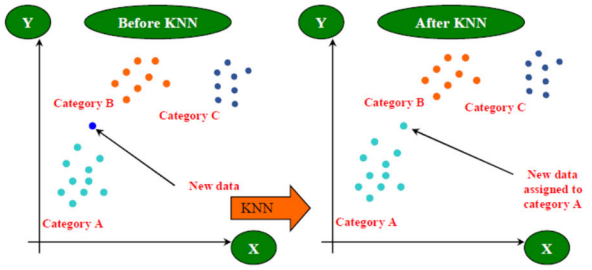


FIGURE 3. *k*-Nearest neighbors.

then it is easy to classify into a well suite category by based on KNN. At the training data just stores the dataset and when it finds new data, KNN classifies that data into a category that is much similar to the new data [21]. KNN is a simple algorithm that stores all available data and classifies new data based on a similarity measure (e.g., distance functions). The Euclidean distance between two points x and y is given by Equation (1) (Figure 3) [22].

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{1}$$

The KNN algorithm:

1- Input the data
2- Initialize the number K of the neighbors
3- Calculate the Euclidean distance of K number of neighbors using Equation (1).
4- Chose the K nearest neighbors.
5- Assign the new data to the nearest category.

C. NAÏVE BAYES

The Naïve Bayes algorithm is a supervised learning probabilistic algorithm that is based on the bayes theorem and is used for classification, estimation, and prediction problems. It is used in disease classification that contains a high-dimensional training dataset. The Naïve Bayes algorithm is one of the simple and effective classifiers in developing fast machine learning models that can build quick prediction models.

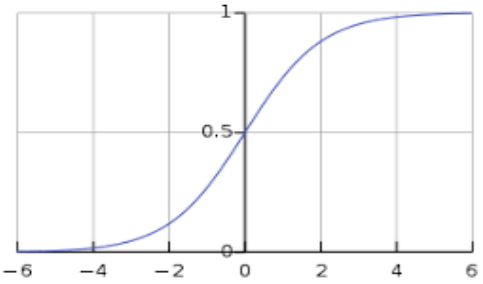


FIGURE 4. Logistic curve.

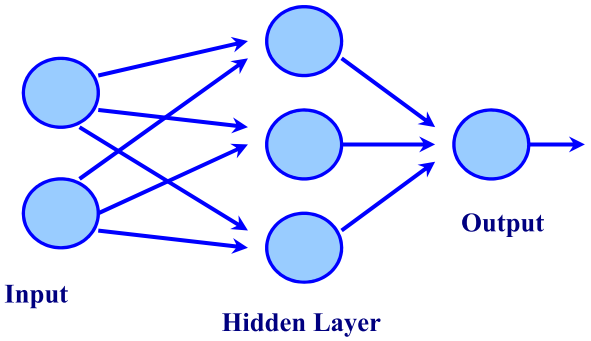


FIGURE 5. Multilayer perceptron.

The Naïve Bayes control several limitations including oversight of complex iterative estimations of the parameter. It can be applied to a large dataset in real-time. Using Bayes theorem posterior probability P(C|X) can be calculated from P(C), P(X), and P(X|C) [23]. Therefore, the Naïve Bayes formula is defined in Equation (2).

$$P(C|X) = (P(X|C)\ P(C))/P(X). \tag{2}$$

where P(C) is class C's probability of being true, P(X) is the predictor's prior probability, P(C|X) is the target class's probability, and P(X|C) is the predictor class's probability [23].

D. LOGISTIC REGRESSION

Logistic regression is one of the most popular supervised learning techniques. It is a model used to predict the categorical dependent variable based on a given set of independent variables. Logistic regression analysis studies the relationship between a categorical dependent variable and a set of independent variables. It gives the probabilistic values that lie between 0 and 1.

Mathematically, a logistic regression model is used to predict $P(Y = 1)$ as a function of X that used for many classification problems like Diabetes prediction and cancer detection, etc. The curve from the logistic function specifies the likelihood of something. When the response is a binary variable and X is numerical, logistic regression fits a logistic curve to the relationship between X and Y. It uses the sigmoid function mapping predicted values to probabilities as Equation 3 [24].

$$S(X) = \frac{1}{1 + e^{-X}} \tag{3}$$

where $S(X)$ is the probability estimate (between 0 and 1), X is the input function, and e is the base of the log

See Figure 4.

E. MULTILAYER PERCEPTRON

Multilayer Perceptron is a parallel distributed information processing structure entailing a compound number of processing elements called nodes. The nodes are interconnected by unidirectional signal channels called connections. Multilayer Perceptron contains one input layer, one output layer, and one or more hidden layers. All layers comprise one or more nodes represented as small circles. The lines between the nodes use to clarify the flow of information from one node to another. The input layer gets signals from external nodes. The outputs from the input layer are moved to the hidden layer by the weighted connection links. In which calculations are performed and then the result is moved to the output layer via weighted links. The outputs from the hidden layer move to the output layer for computations and then produce the result see figure 5 [24].

The working of Multilayer Perceptron in the following steps:

1. Data entered in input layer for processing, which yields a predicted output.
2. Calculate the error value by subtracting the expected outputs from the actual outputs
3. The network uses a back-propagation algorithm to adjust weights.
4. The weights adjusting process starts from weights between output layer nodes and last hidden layer nodes and works backward through the network.
5. When backpropagation is finished, the forwarding process will initiate a new start.
6. The process will be iterated until the error becomes small between the expected and actual output [25].

F. ARTIFICIAL NEURAL NETWORKS

The ANN most commonly machine learning algorithms that used for a wide variety of problems. It is based on a supervised procedure and include three layers: input, hidden, and output. There are a large number of extraordinary kinds of networks, but they are all characterized using the following components: a set of nodes, and connections among nodes. The nodes can be viewed as computational units. They collect input data, which multiple by weights and compute by a mathematical function process to get an output see Figure 6. The connections determine the data go with the flow between nodes. They can be unidirectional, when the data flows solely in one sense, and bidirectional, when the data flows in either sense. The interactions of nodes even though the connections lead to a global behavior of the network, which cannot be

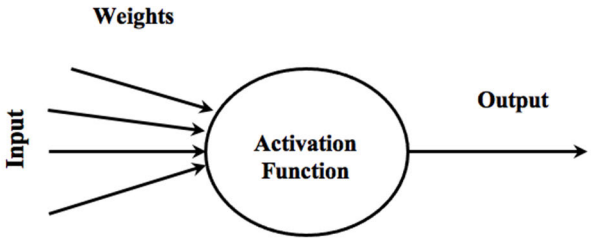


FIGURE 6. Artificial neural networks.

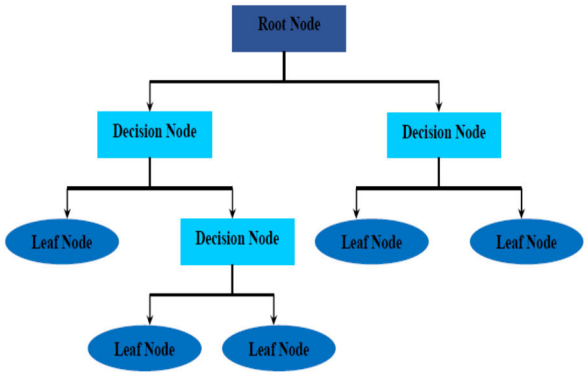


FIGURE 7. Sample decision tree structure.

detected in the factors of the network. This global behavior is stated to be emergent. The fact that the abilities of the community supersede the ones of its elements makes network a very powerful tool [26].

The learning process might be very simple, so it takes place by inputting the dataset to be trained by the network. The information from the input layer is circulated to the hidden layer for information exchange. Then, the output layer will further process the received information to get the results. The predicted values are then compared with the desired values for error calculations [26].

G. DECISION TREE

Decision tree is the most common supervised machine learning technique for solving both classification and regression problems. It is a tree-structured classifier using nodes and internodes. The root and internal nodes represent features of a dataset. The branches represent the decision rules while each leaf node represents the result [26].

There are two nodes in a decision tree: the decision node and the leaf node. The decision node is used to make a decision and has many branches, while the leaf node is the output of those decisions and does not contain any branches see figure 7 [27]. The decision tree provides a powerful technique for the classification and prediction of diabetes.

H. SUPPORT VECTOR CLASSIFIER

Support vector classifier is one of the supervised learning algorithms used for classification, regression, and detection. The goal of the SVC is to build an optimal line or decision

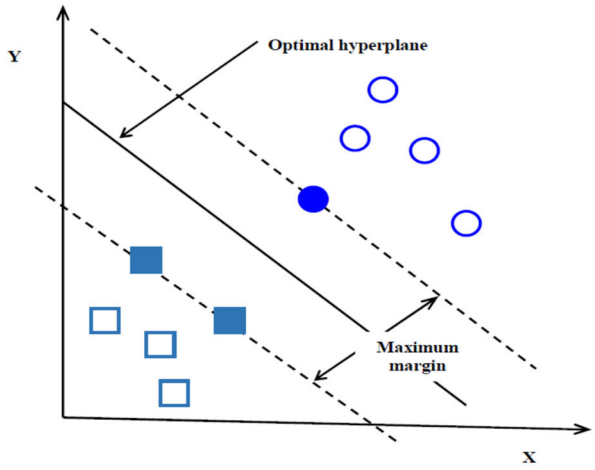


FIGURE 8. Support vector classifier.

boundary that can segregate N-dimensional space into classes and create a plane that has the maximum margin. We can easily put the new data in the correct category. The decision boundary is called a hyperplane. The dimension of the hyperplane depends on the number of features. When the number of features equals two, then the hyperplane is a line. When the number of features is three, then the hyperplane becomes a two-dimensional plane see Figure 8.

For N data points is defined by Equation 4.

$$(\overrightarrow{X_1}, Y_1), \left(\overrightarrow{X_2}, Y_2\right), \ldots, \left(\overrightarrow{X_N}, Y_N\right). \tag{4}$$

where $X1$ is real vector and Y can be 1 or -1 represent the class that $X1$ belongs.

A hyperplane constructed as to maximize the distance between classes y = 1 and y = −1 is defined by Equation 5:

$$\overrightarrow{W} \cdot \overrightarrow{X} = -N. \tag{5}$$

where $\overrightarrow{W}$ is normal vector and $\frac{N}{\left\|\overrightarrow{W}\right\|}$ is offset of hyperplane along $\overrightarrow{W}$ [28].

I. GRADIENT BOOSTING

Gradient boosting is a machine learning technique for regression and classification problems. It is a method for transforming a collection of weak classifiers into one strong classifier. Gradient Boosting model combines the predictions from many decision trees to create the final predictions as in Figure 9.

The objective of Gradient boosting is to define a loss function and minimize it. Loss defined as shown in Equation 6.

$$Loss = \sum \left(y_i - y_i^n\right)^2 \tag{6}$$

where y_i is ith output data and y_i^n is ith prediction. Gradient boosting prediction based on finding the value of loss

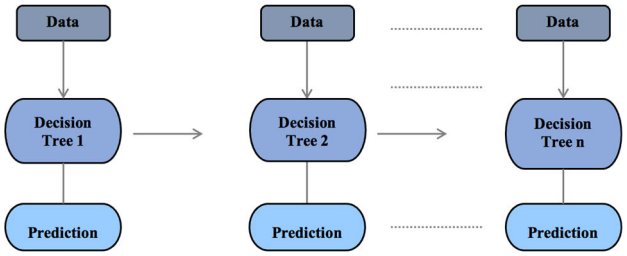


FIGURE 9. Gradient boosting.

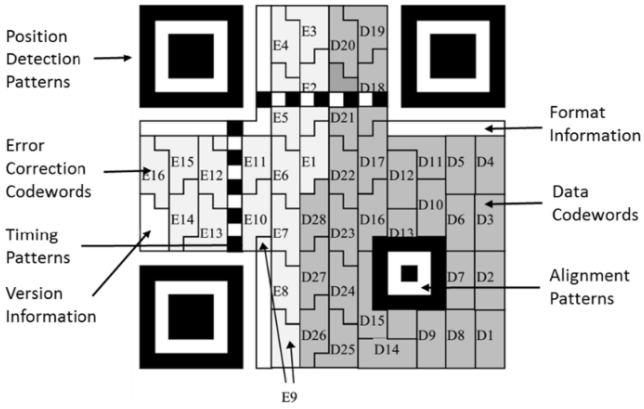


FIGURE 10. QR code structure (version 2 – level L) [31].

function is minimum as Equation 7.

$$y_i^n = y_i^n + a \cdot \sigma \sum \frac{\left(y_i - y_i^n\right)^2}{\sigma y_i^n} \tag{7}$$

where a *is* learning rate and $\sum \left(y_i - y_i^n\right)$ is sum of residuals. The main objective is to make the sum of our residuals close to 0 or minimum and predicted data close to actual data [29].

IV. QR CODE TECHNOLOGY

Denso-Wave, a Japanese company, developed the Quick Response (QR) code, a two-dimensional barcode or matrix code, in 1994 [30], [31]. It is developed to quickly encrypt or decrypt data which is encoding data in both horizontal and vertical directions. While it can encrypt and encrypt data quickly, it cannot store vast amounts of data and is also capable of error-correction [30], [32]. It can also encode a variety of data types. So, in several sectors, QR codes have grown in popularity due to their effective robustness and information density [32]. In our daily lives, we utilize QR codes extensively as a means of transmitting information.

Due of their advantages in quick readability and high capacity, QR codes are widely used in daily lives as a means of transmitting information [31], [33]. They can be used for tracking automotive parts in factories, navigation, advertising, mobile marketing, mobile payments, electronic ticket/coupon, identification, academics, mobile devices, information security, access control, OMR sheet tampering detection [32], [34]. Digital government services

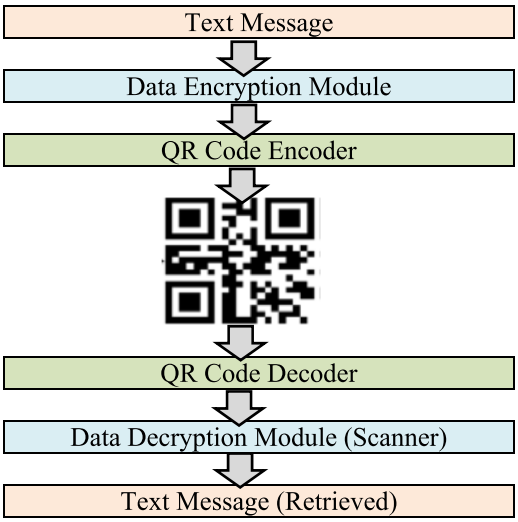


FIGURE 11. Encrypted QR code concept.

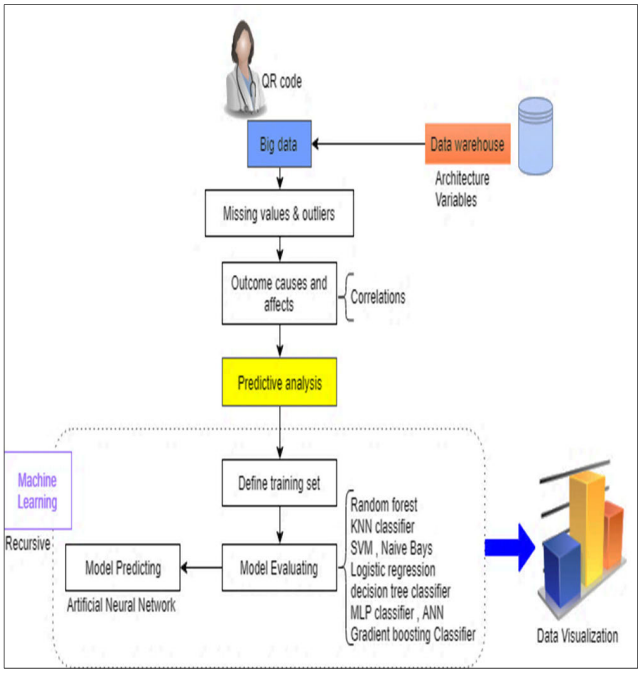
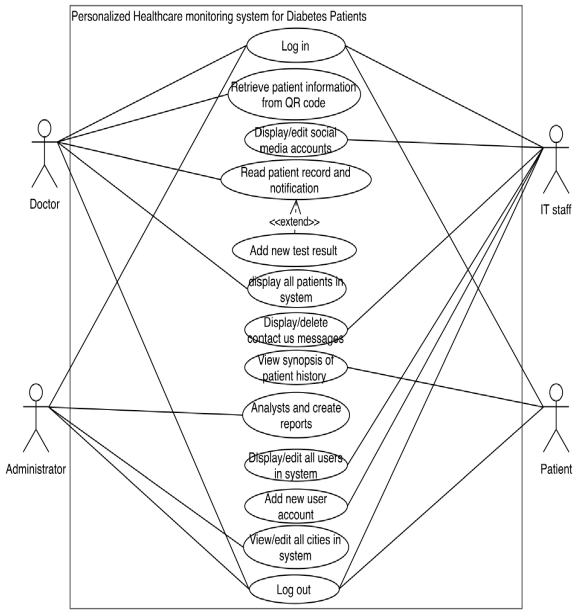


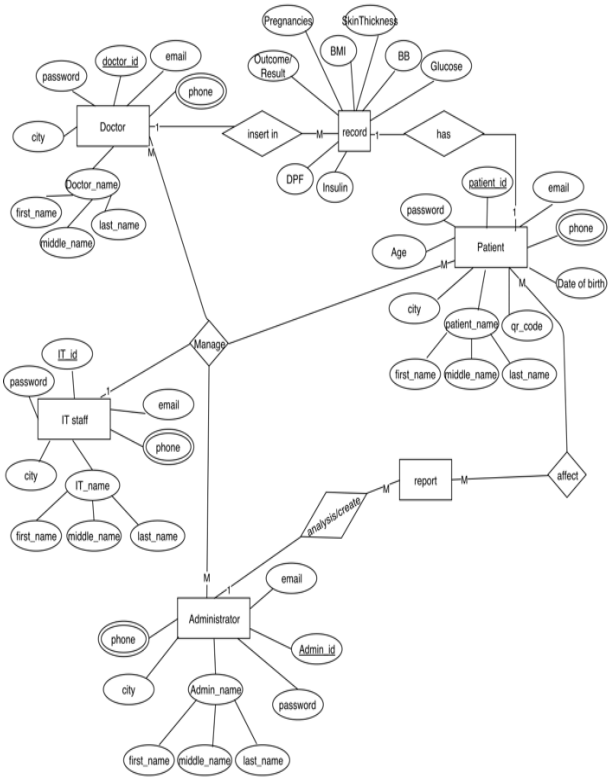
FIGURE 12. Proposed system architecture.

can also employ QR codes to efficiently share important information. Further, QR codes are used to improve users' participation [32]. An effective and secure authentication method for a physical access control system based on QR code technology using a mobile phone and standard equipment was presented by Kao et al. [30]. The researchers save the secret information in QR codes and take OTP's one-time use benefit. According to the research, the proposed solution strengthens the physical access control system's security [30].

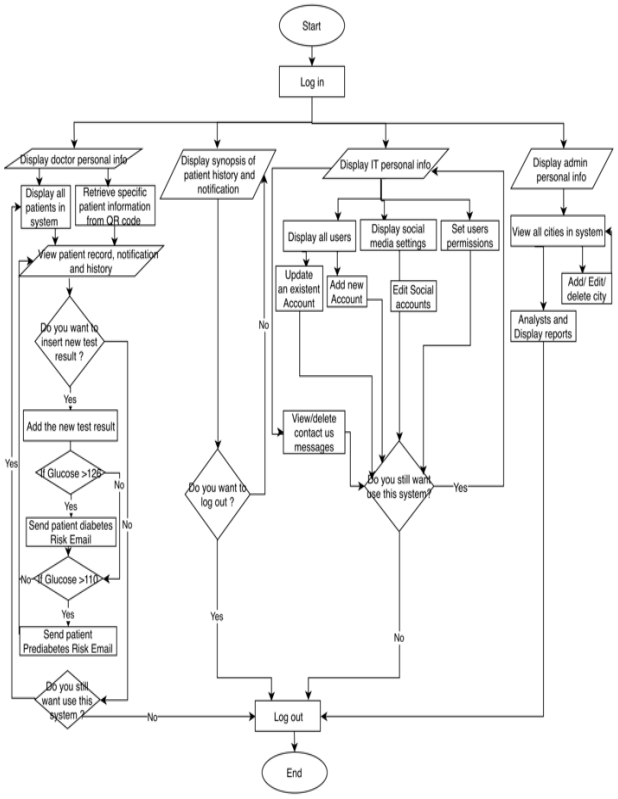
QR code includes a number of white and black modules (squares or dots) where the modules number or the module configuration varies between each version [31], [34]. There



(a) Use Case Diagram.



(b) ERD Diagram.



(c) Data Flow Diagram.

FIGURE 13. **Proposed system use case, ERD, and DFD diagrams.**

are forty different versions of QR codes, and each version has four levels of user-selectable error correction to encapsulate the message of QR code. The QR structure (version 2) with level of error correction (L) is shown in Figure 10. each QR code includes functional patterns like alignment patterns, position detection patterns, timing patterns, version patterns, and error correction codewords [31].

In a system that uses encrypted QR codes, data is first encrypted using cryptographic algorithms before being applied to the QR code generator (encoder) to produce the QR code. The information is obtained by decryption algorithms after this QR Code has been scanned and deciphered using a QR decoder. The description of the encryption and decryption processes of QR code technology is shown in Figure 11 [34].

V. METHODOLOGY

Most diabetics have some knowledge about their health quality or the factors of risk that they will face before diagnosing. In this work, a model is proposed based on machine learning techniques to analyze and predict diabetes (type 2). In this model, we try to improve the performance of the prediction model based on different machine learning algorithms besides making the model adapt to more than one dataset. Moreover, these days the healthcare institutions need a healthcare information platform which fulfill the interaction between medical devices, medical institutions, and patients. Our proposed system (see Figure 12) is developed as a personalized healthcare developing system to support patients to better health situation self-management. The proposed system website is used to monitor, provide feedback, and facilitate communication with medical staff. The main idea behind of our proposed system is to collect patient's health data by medical devices or sensors for diabetic and then sent this data through a wireless network for remoting service platform. Then the data analytics method will be applied to the received data in the healthcare domain. Moreover, the different classifiers are utilized to classify the diabetes patient and predict the diabetic risk. QR card for each patient will include his information that helps the patients and the healthcare institutions, where it is used to support patients to view in their health status: health patterns and future changes prediction. QR card includes the user information such as age, blood, glucose, diabetes pedigree, pregnancies, pressure, BMI, skin thickness, insulin, function, age, outcome, and date.

Our proposed system has a use case diagram (Figure 13(a)) that interacts with four actors (administrator, doctor, IT staff, and patient) and an Entity Relationship Diagram (ERD) (Figure 13(b) that interacts with four actors (administrator, doctor, IT staff, and patient). The Data Flow Diagram (DFD) for the proposed system (Figure 13(c)) displays how data flows through the system and how all users must sign in to use it. The system then uses the credentials to identify and authenticate the users (username and password). After logging in, a user can access the system's services.

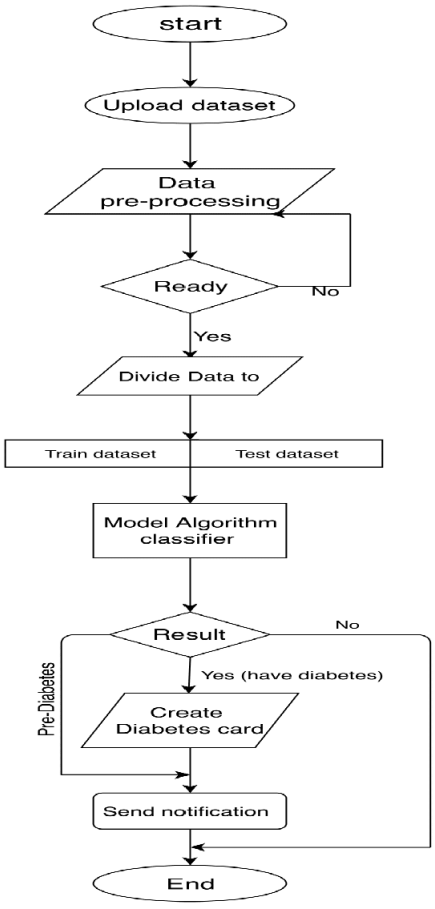


FIGURE 14. Proposed analytical model architecture.

A. PROPOSED DIABETES ANALYTICAL MODEL

In this work, the proposed analytical model (see Figure 14) was developed using Python based on nine ML algorithms to diagnose diabetes in women using 8 important attributes. The proposed model starts with the preprocessing stage where the operations performed on the dataset are replacing the missing data and values normalization. Then processed dataset pass-through feature selection in which sets of attributes are removed from the dataset. After that nine ML algorithms are employed. Finally, for model creation, we apply Split methods and cross-validation. The model output is used to help patients to view in health status their ongoing health patterns and future adjustments.

1) DATASETS

In this work, the proposed analytical models are evaluated based on two different datasets that are Synthetic dataset (SD) and the PIMA diabetes dataset (PIDD) [35]. SD dataset was generated by merging more than datasets specifically from the Kaggle data repository and includes medical details of 7691 instances for female and male with seven attributes. It is prepared by removing the noisy and inconsistent data and then a new column (height) was computed by the

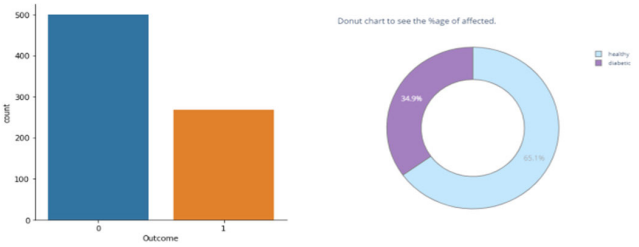


FIGURE 15. Distribution of healthy and diabetes people.

TABLE 1. Dataset description.

No. of Attribute Database.	No. of Instances
8	768

TABLE 2. Dataset sample.

Age	Glucose	Pregnancies	Skin Thickness	Blood Pressure	BMI	Insulin	Diabetes Pedigree Function	Outcome
50	148	6	35	72	33.6	0	0.627	1
31	85	1	29	66	26.6	0	0.351	0
32	183	8	0	64	23.3	0	0.672	1
21	89	1	23	66	28.1	94	0.167	0
33	137	0	35	40	43.1	168	2.288	1

average height for the women and men to calculate the BMI which is the main factor for diabetes and consequently a new column (BMI) was added using the formula (weight / height * height) and the (weight, temp) columns were removed. Then a new column (risk) was added using this rule (=IF(AND(name of glucose column >=126; name of glucose column<500);"2";IF(AND(name of glucose column >=100; name of glucose column<126);"1";"0")) [36]. PIDD was taken from the UCI data repository and includes medical details of 768 female patient instances. In addition, the dataset includes 8 attributes of numeric value. Where the value of one class '0' is handled to test negative for diabetes and the value of another class '1' is handled to test positive for diabetes. See description of Dataset in Table 1. The dataset sample is indicated in Table 2. There are 500 healthy people in the dataset, and 268 people with diabetes as shown in Figure 15.

Figure 16 shows a correlation heatmap of various attributes; the correlation coefficient between 'Glucose' and 'Outcome' is 0.47. The significant correlation between 'Age' and 'Pregnancies' is 0.54, which is self-explanatory therefore as a woman's age increases; the pregnancy rate tends to increase. Figure 17 depicts the relationships between all

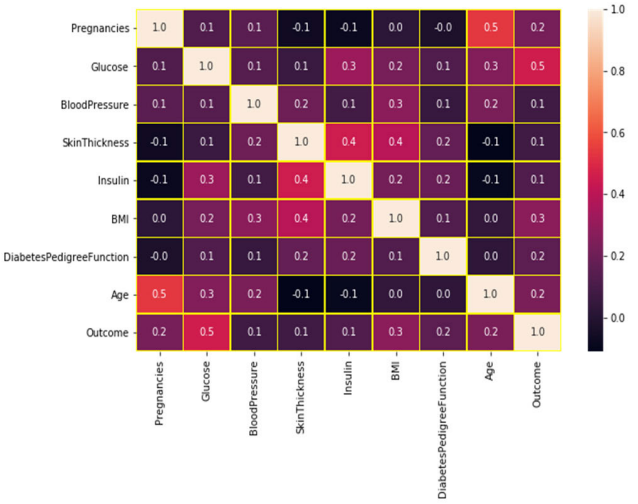


FIGURE 16. Correlation heatmap.

FIGURE 17. Relationship between all attributes.

attributes concerning one another using the *pairplot* method in *seaborn*. The hue property could be used to distinguish between distinct classes of each feature. The 'Outcome' of the dataset is the hue employed. The figure demonstrates an increase in the risk of diabetes as glucose levels rise, where the blue color indicates $Outcome = 0$ and $Outcome = 1$ for the orange color. Figure 18 displays the significance of features, where skin "*Thickness*" is the least important attribute and "*Glucose*," "*DiabetesPedigreeFunction*," and "*BMI*" are the most important.

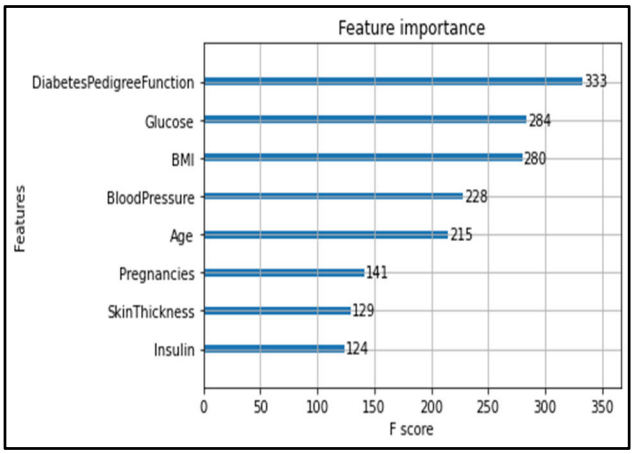


FIGURE 18. Significance of the features.

2) DATA CLEANING AND TRANSFORMATION

Using Data Visualization in python, the attributes can be sorted in different ways to view patterns and values. Some attributes have values of zero despite the fact that it can't be an appropriate value (missing value). Those invalid values should be replaced with median as central tendency and substitute when value is zero for the attributes Plasma-Glucose, Diastolic BP, and BMI. So in data transformation, the masks were created for invalid values of each column, then all invalid values were replaced with zeros. As the distribution of all data is either normal or skewed, the median would be calculated as a central tendency and substitute when the value is zero. All attributes haven't transformed because for some of them can be zero (like "Number of times pregnant").

3) DATA PREPARATION

Data preparation aims to ensure that the data prepared for the analysis is accurate and consistent, so BI results and Analytics applications will be valid. Data is often generated with missing values, errors, or other errors. In addition, datasets that are stored in separate files or databases often contain different formats that need to be reconciled. Debugging, checking, and joining datasets are a big part of the data set-up process. In large data applications, data preparation is a highly automated task, as IT staff or data analysts may take a lot of time to correct each manual field in each file to be used for analysis. Machine learning algorithms can check data fields, filling empty values automatically, or renaming specific fields to ensure consistency when data files are joined. In our proposed system, the input data is processed to improve the system.

Regarding PIDD, the ranges of numerical attributes differ from each other, such as the range of BMI (18.2 - 67.1) and Plasma-Glucose (44.0 - 199.0). This type of variation can lead to a trend variable with a greater range to have an unnecessary impact on the results. Also, the dataset standardize the scale of effect for each variable by normalizing their numerical variable [37]. Standard Scaler is used in this research. It assumes data is normally distributed within each feature and will scale them such that the distribution is now centered around 0, with a standard deviation of 1. The mean and standard deviation are calculated for the feature and then the feature is scaled based on Equation (8):

$$\frac{xi - mean(x)}{stdev(x)}. \tag{8}$$

The process of data preparation stage is shown in

Algorithm 1 Data preparation

Input : *PIDD dataset*
Output: prepared dataset

Start
Step 1: load the PID dataset
Step 2: remove the noisy and inconsistent data
Step 3: add a new column (risk) using this function (=IF (AND (name of glucose column >=126; name of glucose column<500);"2"; IF (AND(name of glucose column >=100; name of glucose column<126);"1";"0"))
Step 4: return prepared dataset
End

4) TECHNIQUES FOR CROSS-VALIDATION AND DATASET SPLITTING

The dataset was split into a training dataset of around 80% and a test dataset of around 20% as it is a very important step for supervised machine learning models. The first part was used to train the model, and then the trained model was used to make predictions. K-folds cross-validation was conducted based on the dataset. For each K experiment, (K-1) numbers are used for training and the remaining one for testing. K-Folds with 5 shuffled folds were created to validate the dataset used to design the Success rate for detection.

5) CLASSIFICATION PHASE

In the classification phase, each pattern of the records that occurred from the preceding stage will be allowed to whether wholesome or diabetic. See the steps of the proposed classification approach (Algorithm2) as follows:

Algorithm 2 Classification

Input: Prepared dataset
Output: Classification results

start
Step1: Load the input dataset
Step2: Divide the input data into number of subsets
Step3: Run classifier on each subsets
Step4: Apply voting majority on the collected results
Step6: Return (classification result)
End

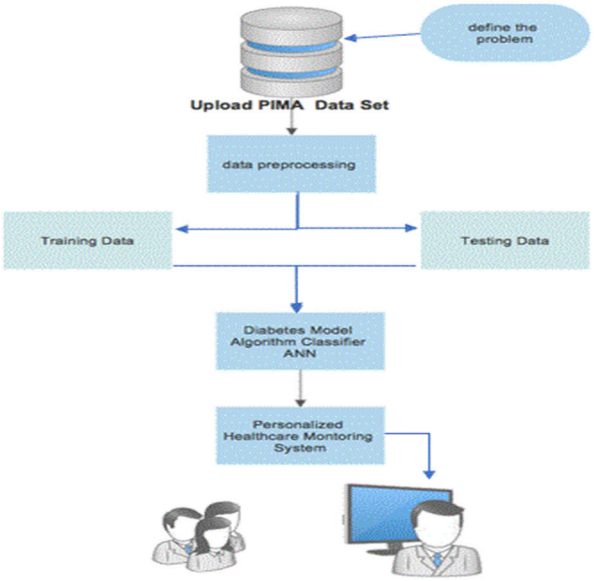


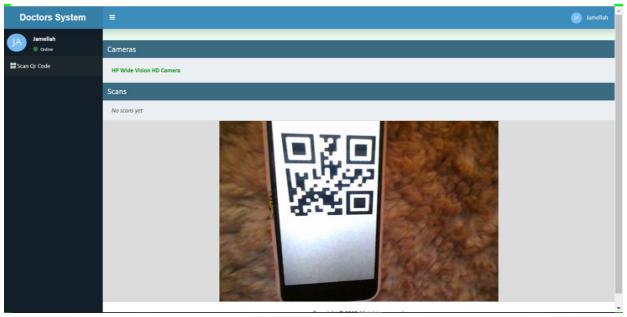
FIGURE 19. Workflow of implementation.

B. PROPOSED WEBSITE IMPLEMENTATION

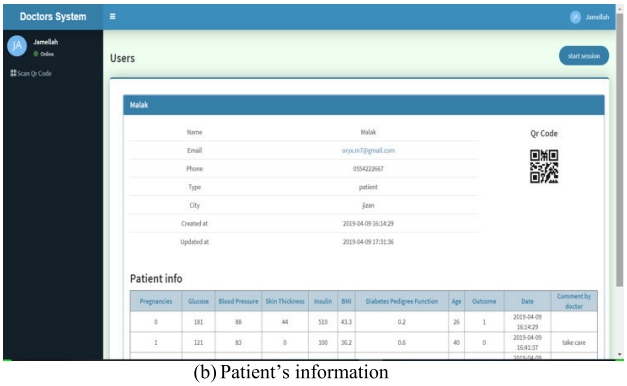
Website is the most frequently used technology in diabetes care. The main features of the proposed website are the QR card system, communication and data visualization system, and Sending web-alert emails to patients to alert the patient of diabetic risks. The web development tool that was used is the Sublime Text, using HTML, PHP, CSS, and JavaScript; the database used to store patients' and doctors' information is MySQL. Also, many free and open-source frameworks are utilized such as BootStrap for an open-source front-end web framework, Laravrel PHP is an open-source PHP web framework and Vue.js is an open-source JavaScript framework. The Implementation workflow is indicated in figure 19.

QR card system (see Figure 20), we can get QR code-based medical identification alerts and an in-hospital patient identification system by the proposed model. Each patient in the system is took a unique QR code tag; to assist with medical identification alerts. The most important feature in our system is the QR card that collects the user information such as age, glucose, insulin, blood pressure, BMI, diabetes pedigree function, skin thickness, pregnancies, outcome, and date. This QR card is owned by the patient to move easily between the health centers. QR card is used to recognize the person and get his/her necessary information. So, the doctor can access patient information easily. Figure 20 indicates the process of the proposed QR code system which includes scanning the patient's QR code and then the patient information will be displayed.

Regarding the communication and data visualization system, all users (doctors, IT staff, admin, and patients) can access the system. When the QR code is scanned, all patient information appears as shown in Figure 20. Also, by choosing the ''User'' and then viewing the patient, the doctor will be



(a) Scanning QR code.



(b) Patient's information

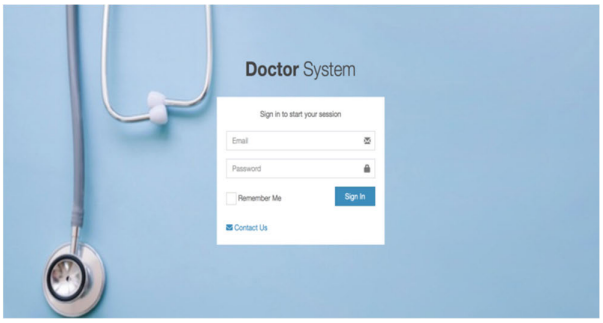
```
</div>
@if($user->type == "patient")
<div class="col-md-3 col-lg-3 " align="center">
    <h4>Qr Code</h4>
    {{(QRCode::text($user->id)->svg())}}
</div>
@endif
<div class="table-responsive col-md-12">
@if(isset($patient_info[0]))
        <h3>Patient info</h3>
        <table class="table-responsive" border="1"  >
        <tr style="background-color: #bcd1dc;font-weight: bold;color: #367fa9;">
            <td >Pregnancies</td>
            <td >Glucose</td>
            <td style="width: 10%">Blood Pressure</td>
            <td style="width: 10%">Skin Thickness</td>
            <td >Insulin</td>
            <td >BMI</td>
            <td style="width: 20%">Diabetes Pedigree Function</td>
            <td >Age</td>
            <td >Outcome</td>
            <td style="width: 10%" >Date</td>
            <td style="width: 10%">Comment by doctor</td>

        </tr>
        @foreach($patient_info as $patient)
            <tr>
                <td>{!! $patient->pregnancies!!}</td>
                <td>{!! $patient->glucose!!}</td>
                <td>{!! $patient->bloodPressure!!}</td>
```

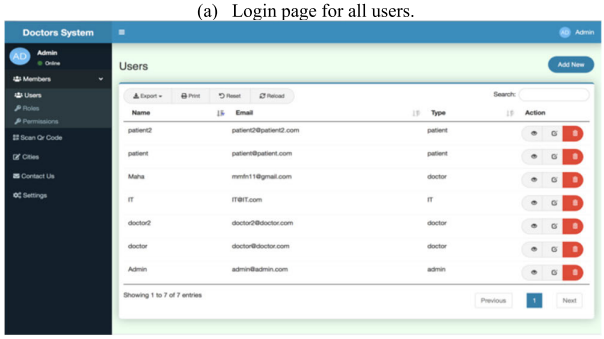
(c) QR source code

FIGURE 20. QR card system.

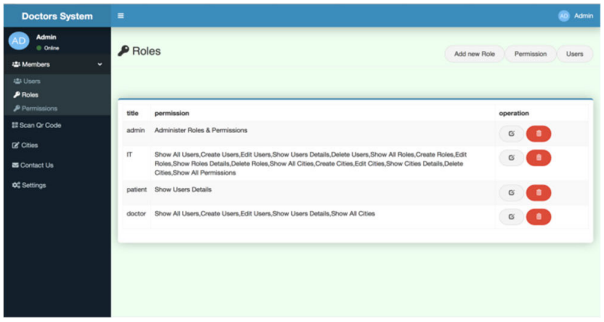
able to see the patient's information via the QR code, then the table containing the patient's information will be displayed to the doctor. Although the most important and helpful element is the doctor's observations, and this will be displayed to any doctor even in other healthcare institutions when needed. The web-based application also includes web alert emails to the patients which provide messages and information about diabetic patients' risk of readmission. The message content will be sent if the glucose of the patient is >122, where the blood glucose levels were categorized into hyperglycemic (>126 mg/dl), normoglycemic (70-126 mg/dl), and hypoglycemic (<70 mg/dl). The web-based application is aimed at the response time to any patient's concern within a few hours



(a) Login page for all users.

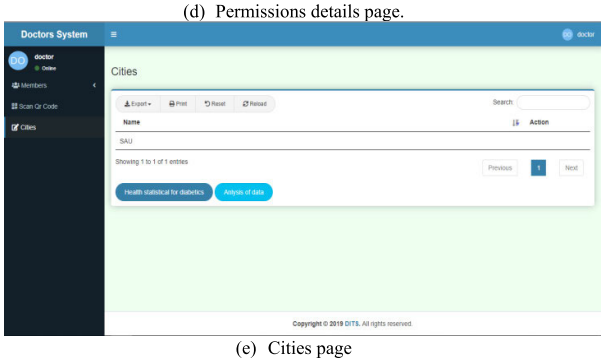


(b) All Users account from admin page.



(c) Roles page.

(d) Permissions details page.



(e) Cities page

FIGURE 21. Main input and output pages in proposed system.

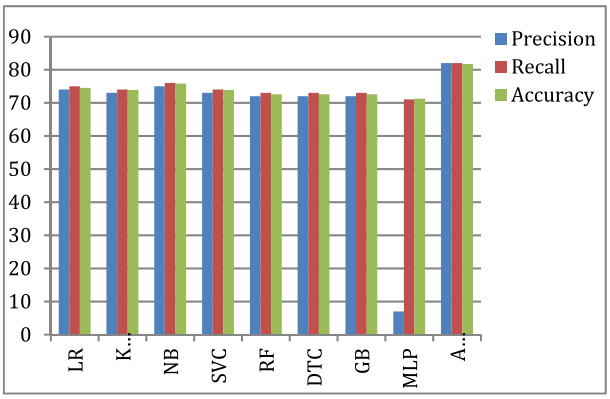


FIGURE 22. Performance of classification models (PIDD).

and continuously any patient can follow up on his/her condition. Subsequently, that's will save time and avoids the risk.

The database is a separate part and aims to stores the collecting data. DataBase on MySQL was established, including a systematic breakdown of how it works with different programming languages, and the important advantages are increasing the performance of applications, fast, portable, reusable, transparent, and also secure. The disadvantages is memory usage increased, restricted for complex business logic also difficult to debug and maintain. First, the user table was designed including (id, name, email, email_verified_at, password, phone, type, remember_token, created_at, update_at, deleted_at). On the admin structure includes a set of variables (Name, Type, Collection, Attribute, Null, Default, Comments, Extra, Action), which is the primary key is (id), and the foreign key is (email, city_id). This Admin attribute is applied to the rest of the DB tables such as patient, IT and doctor. In the roles table, there is a list of variables which is (id, name, guard_name, created_at, updated_at). Figure 21 shows the main input and output webpages in proposed system.

VI. RESULTS AND DISCUSSION

A. DIFFERENT FEATURES OF THE PROPOSED SYSTEM

The developed platform via which a patient's diabetic data is gathered, processed using machine learning methods to decide whether the patient has diabetes or not, and then a card is made for the patient to help him navigate between medical institutions and turn it into an electronic health record. The system has a dashboard that shows the number of patients in Saudi Arabian cities. Additionally, it has statistical charts that show the classification of patients' conditions (normal, Pre-Diabetes, Diabetes). The developed system has different features to achieve further functions (see Table 3) in addition to the data analysis and prediction model construction of diabetes diseases.

B. DIABETES CLASSIFICATION MODELS

This section summarizes the findings of the experiments from the evaluation of the analytical model based on SD and PIDD

TABLE 3. Different features of the proposed system.

Users	Features	Functionality
Doctor	View list of registered patients.	displaying a table that has a list of patients along with allowing a doctor to view patients details.
	Search for specific patient	allowing a doctor to search for patient by name.
	Export /print patients' data	allowing a doctor to print dataset from system patients or release it in different format such as excel or csv.
	Reset/ reload patients' data	allowing a doctor to reset data from system patients or reload it from outside source.
	View patient information and vital information	displaying a page that has full patient information such as name, email and city …etc. Also, vital information such as glucose, insulin and bmi …etc.
	Edit patient information and vital information	allowing a doctor to edit vital information such as glucose, insulin and bmi …etc.
	Scan QR code	allowing a doctor to scan QR code and get patient information.
	Predicting diabetes mellitus for patient	allowing a doctor to predict if patient has diabetes or not depending on patient vital information.
	View doctor information	allowing a doctor to view her/his information.
Patient	Login successfully	The login functionality where only authorized system patients can successfully log into their accounts
	View Patient information and Vital information	displaying a page that has patient personal information such as name, email and phone …etc. In addition, Vital information such as Glucose, Insulin and BMI …etc.
	Logout successfully	The logout functionality where patients can successfully log out of their accounts
IT user	Register new user	The functionality of successfully registering a new user such as doctor, patient, admin or even another Its user into the database.
	Edit existing user	displaying a page that has an information of specific user.
	Delete existing user	The functionality of successfully removing user from the database.
	View list of registered users.	displaying details of users along with allowing an IT to view each user profile
	Search for specific users	allowing an IT to search for user by name.
	Export / Print users' data table	allowing an IT to print dataset from system users or release it in different format such as excel or csv.
	Reset / Reload users' data table	allowing an IT to reset data from system users or reload it from outside source.
	Set permissions for all users	allowing an IT to edit permissions for all users, such as patients can only read their information.
	View / Edit Social media accounts	allowing an IT to View and edit Social media accounts, such as Facebook, Twitter, and Instagram.
	View users' messages or questions from contact us page	displaying a table that has list of people with their emails and phone numbers coming from contact us page along with allowing to view their the messages, questions, or suggestions.
	Delete users' messages or questions from contact us page	allowing an IT to delete the messages, questions, or suggestions that came from contact us page
Admin	Visualize Health Statistical for diabetics	Visualizing Statistical reports according to the number of patients with diabetes with non-patients.
	Visualize number of patient's risk status	Visualizing Statistical reports depending on number of patients' risk status
	Export / Print cities data table	allowing an admin to print cities data from system or release it in different format such as excel or csv.
	Reset / Reload cities data table	allowing an admin to reset cities data from system or reload it from outside source.
	Viewing cities list	correctly displaying a page that has a list of cities.
	Add new city successfully	allowing an admin to add a new city.
	Edit existing city	allowing an admin to edit city name.
	Delete existing city	successfully removing city from the database.
	View all user's information	allowing an admin to displaying a page that has all users' information.
	Search for specific city	allowing an admin to search for city by name.

TABLE 4. Performance measurements (PIDD).

Model	Precision	Recall	Accuracy
LR	74	75	74.5098
KNN	73	74	73.8562
NB	75	76	75.817
SVC	73	74	73.8562
RF	72	73	72.549
DTC	72	73	72.549
GB	72	73	72.549
MLP	70	71	71.2418
ANN	82	82	81.6993

datasets. Each dataset is split into training and testing subsets. In terms of accuracy, recall, and precision, the classification models' performance was examined based on the method of K-fold cross-validation. The evaluation findings are analyzed and compared.

The performance of DTC, RF, LR, ANN, NB, KNN, SVC, GB, and MLP models based on PIDD dataset is shown in Table 4 and Figure 22. The findings are found that the ANN approach has better performance than other approaches. By applying the ANN method, four layers were added with the ReLU activation function with 50 units in each layer. After the data analysis, the classification of patients' data was analyzed into three types that are (0-Normal, 1- Pre-Diabetes, and 2- Diabetes) as shown in Figure 23. Where a new column (Risk) was added using the formula (=IF(AND(name of glucose column >=126; the name of glucose column<500);"2"; IF(AND(name of glucose column >=100; the name of glucose column<126); "1";"0")) [36].

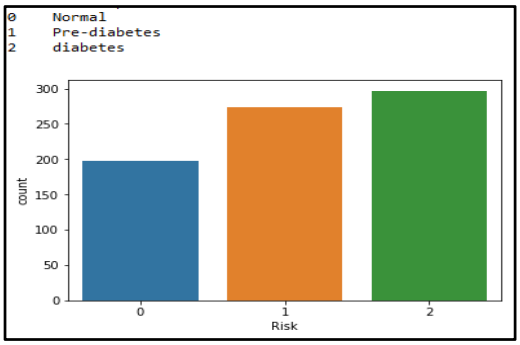


FIGURE 23. Diabetic patients' risk of readmission.

TABLE 5. Performance measurements (SD).

Model	Precision	Recall	Accuracy
LR	86	84	81.77
KNN	78	75	72.27
NB	85	82	80.07
SVC	84	82	82.68
RF	88	87	86.85
DTC	90	85	87.37
GB	88	86	87.49
MLP	83	82	83.46
ANN	87	84	83.72

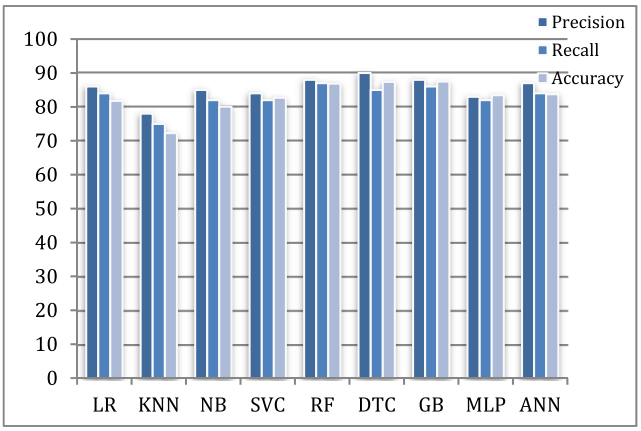


FIGURE 24. Performance of classification models (SD).

Table 5 and Figure 24 represent the performance of classification algorithms based on SD dataset that includes 7691 instances. The results are 81.7% using LR, 72.2% using KNN Algorithm, 80.07% using NB, 82.8% using SVM 86.85% using RF, 87.36% using DTC similar result with GB, 83.45% using MLP, 83% using ANN. So the performance of DTC and GB can predict diabetes with more accuracy as compared to other algorithms. Table 6 indicates a comparison between the proposed models and some existing classification models.

C. DIABETES MONITORING SYSTEM

This work aims to develop an analytical predictive model based on machine learning techniques and a web-based personalized diabetes monitoring system to predict type-2 diabetes. The whole history of a patient can be captured and ready for future analysis using machine learning algorithms by continuously monitoring the patient's vital data. The proposed monitoring system should be scalable enough to handle an increasing volume of patient data without loss of performance. The proposed system keeps track of the users' various health-related activities. It can also be considered a healthcare records platform that facilitates communication between patients and medical organizations. The system's major goal is to gather patient data and present it to specialists so that they can give the appropriate responsible decision. The code can be scanned with any QR Code reader to obtain emergency contact health information, but when the code is scanned with QR Code Identity, additional information is gathered and revealed (see Figure 20).

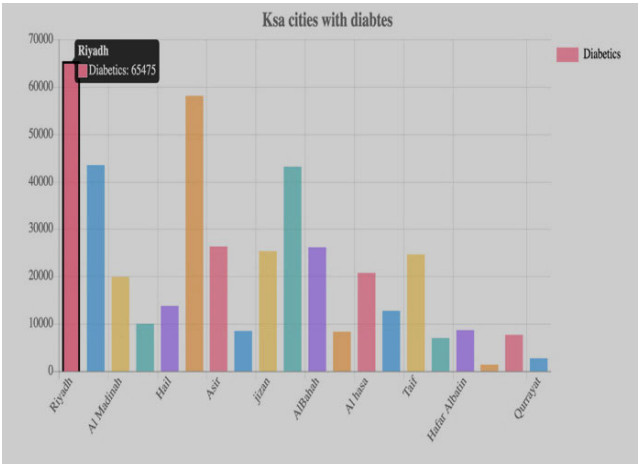


FIGURE 25. Number of diabetics in Saudi Arabia.

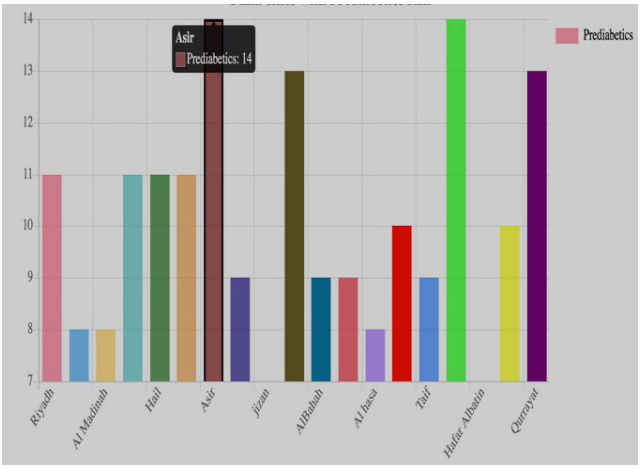


FIGURE 26. Number of who are at risk in Saudi Arabia.

D. CITIES DASHBOARD BASED ON DATA VISUALIZATION

Data Visualization helps people to understand the data by placing it in visual information. Powerful data visualization and dashboards can be built by utilizing JavaScript data visualization and many Libraries in JavaScript. So, by using web technologies (chart.js library), different types of Charts

TABLE 6. A comparison between the proposed and existing models.

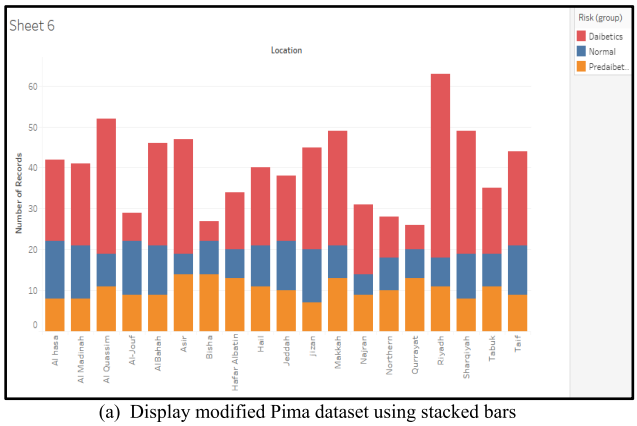
	Proposed System	Article [36]	Article [6]	Article [38]
Problem	If there is an emergency and we need to read the previous reading of sugar for the patient to take the appropriate action.	Diabetes is a significant expense on healthcare systems, mainly in developing nations.	In hospitals that serve a huge population, patient readmission is common.	Difficulty on Data collection process in real time, need to personalized analysis of big data from different Sources and need to continuous suggestions of diabetes.
Users	Doctors, patient, IT, admin.	Patient and medical team.	Patient, medical team, big data analysis.	Patients, relatives, friends, personal health advisors, and doctors.
Solution	A patient's card provides personal information about the patient such as name, age, insulin readings, glucose, pressure measurement and body mass index to help the patient move between healthcare institutions.	Using a real-time data processing and BLE-based sensor device, the researchers presented a personal healthcare monitoring system. Website for medical team and application for Patient.	The researchers developed a technique to identify risk factors that may lead to readmission in diabetes individuals.	The researcher developed a 5G-Smart Diabetes system that would provide patients with diabetes with comprehensive sensing and analysis. They also demonstrate how to share data and create a 5G Smart Diabetic test bed.
Dataset	PIDD Database	CGM Dataset.	The dataset was taken from the UCI Machine Learning Databases Repository. It was chosen from the National Institutes of Diabetes, Digestive, and Kidney Diseases' bigger data source.	The dataset includes 12,366 persons and 757,732 data elements. It is from a hospital in Hubei Province, China.
Algorithm	▪ Decision Tree ▪ Logistic Regression ▪ Support Vector Clustering (SVC) ▪ Naive Bayes ▪ KNN ▪ Random Forest ▪ ANN ▪ Gradient Boosting	▪ Machine Learning Approaches ▪ MLP	▪ Decision Tree ▪ Logistic Regression ▪ KNN. ▪ SVM	▪ Decision Tree ▪ SVM ▪ ANN
Performance based on Accuracy	▪ LR = 74.5098% ▪ KNN=73.8562% ▪ NB=75.817% ▪ SVC=73.8562% ▪ RF=72.549% ▪ DTC=72.549% ▪ GB=72.549% ▪ MLP=71.2418% ▪ ANN=81.6993%	▪ RF =73.046% ▪ NB=76.6927% ▪ SVM=76.562% ▪ Logistic Regression=76.0417% ▪ MLP=77.083%	▪ Feature: correlation value ▪ Plasma glucose: 0.4665814 ▪ Pregnant: 0.2218982 ▪ Blood pressure: 0.06506836 ▪ Insulin: 0.130548 ▪ Skin: 0.07475223 ▪ Body Mass Index: 0.2926947 ▪ Age: 0.23835 ▪ Pedigree Function: 0.1738441	▪ Decision tree about 91% ▪ ANN about 80%, ▪ SVM about 92%

can be drawn based on HTML5 canvas element. The number of diabetics and the number of who are at risk in Saudi Arabia was needed to be displayed, described and illustrated the data. So, the proposed website based on the data visualization has Cities webpage as an interactive dashboard (see Figures 25 and 26) to help the website Healthcare users to view the diabetics' information in Saudi Arabia for analyzing data and gleaning insight and consequently make decisions.

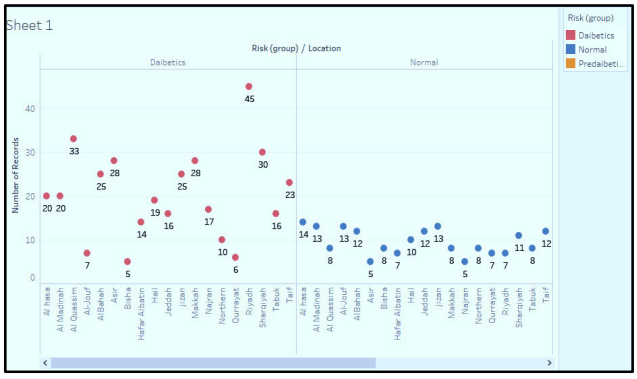
For productive data exploration, the Tableau desktop application was used to generate Statistical reports (see Figure 27) depending on the Modified PIMA dataset and Saudi Arabia diabetes Statistics in the year (1432H (Hijri) - 2010) by using different visualization types (such as bar charts, line charts, histograms, and so on).

VII. CONCLUSION AND FUTURE WORK

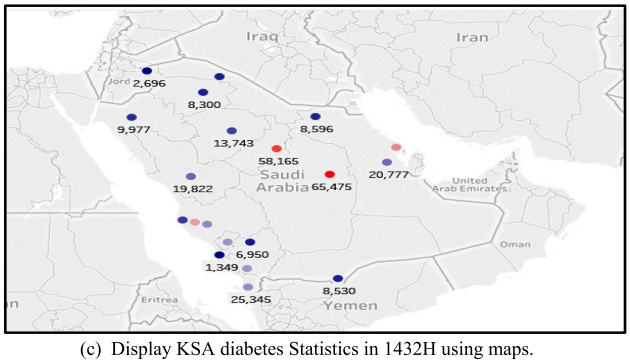
This work aims to develop an analytical predictive model based on machine learning techniques and a web-based personalized diabetes monitoring system to predict type-2 diabetes. The data of a patient will be collected and ready for subsequent analysis based on machine learning models by constantly monitoring the users' vital information. This system is designed to overcome some problems in the health care of diabetics. Predictive models were proposed for diabetic data analysis in big data and IOT. Nine prediction models are compared LR, KNN, NB, SVC, RF, DTC, GB, MLP, and ANN for predicting diabetes using the most important attributes. The development mechanism in the system goes through stages; First, by designing the site through which the



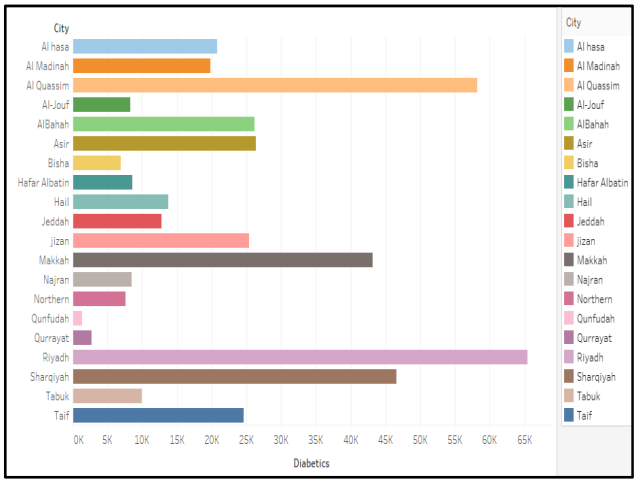
(a) Display modified Pima dataset using stacked bars



(b) Display modified Pima dataset using Side-by-Side circles.



(c) Display KSA diabetes Statistics in 1432H using maps.



(d) Display KSA diabetes Statistics in 1432H using horizontal bars

FIGURE 27. Statistical reports based on Tableau desktop application.

diabetic data for a patient is recorded. Then analyze the data based on different machine learning models using Python to determine if the patient is having diabetes or not, after that, a QR card is created for diabetes patients to help them move between healthcare centers and make it an electronic health record for the patient. The system has dashboard graphs displaying the number of patients in Saudi Arabia cities. It also contains visualized graphs that include more detailed classifications for patients' states (Normal, Pre-Diabetes, and Diabetes).

The proposed monitoring system will be modified in future work to accommodate any other chronic diseases with appropriate datasets. Furthermore, the study can be expanded and enhanced to automate diabetes analysis using various machine learning algorithms. The proposed model, on the other hand, can be enhanced to handle images of diabetic patients as well as audio data collected from them via oral interactions and other audio devices. It is also intended to introduce additional features that will make the system more efficient and provide more services to healthcare institutions. These capabilities include the ability to construct QR card for other diseases and convert it to an electronic file, as well as the insertion of the patient's medications. Creating alerts and using them to remind patients of deadlines, provide suggestions and nutrition programs, and submit the results of the patient's medical testing through the laboratory analyst.

ACKNOWLEDGMENT

The authors would like to thank all the participants involved into this work specially to H. Alruwaili, A. Alanazii, A. Alqahtani, J. Almutairi, L. Alkhulifi, and M. Aldossari.

REFERENCES

[1] WHO Expert Committee on Diabetes Mellitus and World Health Organization. (1980). *WHO Expert Committee on Diabetes Mellitus [Meeting Held in Geneva From 25 September to 1 October 1979]: Second Report*. Accessed: Jun. 4, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/41399

[2] P. Rahimloo and A. Jafarian, "Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them," *Bull. Société Royale Sci. Liège*, vol. 85, pp. 1148–1164, Jan. 2016, doi: 10.25518/0037-9565.5938.

[3] C. Bhatt, N. Dey, and A. S. Ashour, Eds., *Internet of Things and Big Data Technologies for Next Generation Healthcare*, vol. 23. Cham, Switzerland: Springer, 2017, doi: . doi: 10.1007/978-3-319-49736-5.

[4] K. Venkatachalam, P. Prabu, A. S. Alluhaidan, S. Hubálovský, and P. Trojovský, "Deep belief neural network for 5G diabetes monitoring in big data on edge IoT," *Mobile Netw. Appl.*, vol. 27, no. 3, pp. 1060–1069, Jun. 2022, doi: 10.1007/s11036-021-01861-y.

[5] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, 1997.

[6] S. Salian and D. G. Harisekaran. (2015). *Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients*. Accessed: Jun. 2, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Big-Data-Analytics-Predicting-Risk-of-Readmissions-Salian-Harisekaran/ab3c6a823e7f8bcee87f602a1b9d451560fac408

[7] N. M. S. Kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive methodology for diabetic data analysis in big data," *Proc. Comput. Sci.*, vol. 50, pp. 203–208, May 2015, doi: 10.1016/j.procs.2015.04.069.

[8] A. Iyer and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," *Int. J. Data Mining Knowl. Manage. Process*, vol. 5, no. 1, pp. 1–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101.

[9] S. Rai, "Analysis of diabetic data set using hive and R," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 4, pp. 626–629, Jul. 2014.

[10] J. L. Breault, C. R. Goodall, and P. J. Fos, "Data mining a diabetic data warehouse," *Artif. Intell. Med.*, vol. 26, no. 1, pp. 37–54, Sep. 2002, doi: 10.1016/S0933-3657(02)00051-9.

[11] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Jul. 2020, doi: 10.1038/s41598-020-68771-z.

[12] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," *Int. J. Eng. Res. Appl.*, vol. 3, no. 2, pp. 1797–1801, Apr. 2013.

[13] G. Krishnaveni and T. Sudha, "A novel technique to predict diabetic disease using data mining—Classification techniques," *Int. J. Adv. Sci. Technol., Eng. Manag. Sci.*, vol. 3, pp. 1–7, Mar. 2017.

[14] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Jun. 2020, doi: 10.1007/s40200-020-00520-5.

[15] H. Zhou, R. Myrzashova, and R. Zheng, "Diabetes prediction model based on an enhanced deep neural network," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, p. 148, Jul. 2020, doi: 10.1186/s13638-020-01765-7.

[16] A. Alharbi and M. Alghahtani, "Using genetic algorithm and elm neural networks for feature extraction and classification of type 2-diabetes mellitus," *Appl. Artif. Intell.*, vol. 33, pp. 1–18, Dec. 2018, doi: 10.1080/08839514.2018.1560545.

[17] P. B. Harleen, "A prediction technique in data mining for diabetes mellitus," *J. Manage. Sci. Technol.*, vol. 4, no. 1, pp. 1–12, 2016.

[18] D. V. Dimitrov, "Medical Internet of Things and big data in healthcare," *Healthcare Inform. Res.*, vol. 22, no. 3, pp. 156–163, Jul. 2016, doi: 10.4258/hir.2016.22.3.156.

[19] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Making*, vol. 11, no. 1, p. 51, 2011, doi: 10.1186/1472-6947-11-51.

[20] P. Sittidech and N. Nai-Arun, "Random forest analysis on diabetes complication data," in *Proc. IASTED Int. Conf.*, 2014, pp. 315–320, doi: 10.2316/P.2014.818-047.

[21] K. Saxena, Z. Khan, and S. Singh. (2014). *Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm*. Accessed: Jun. 4, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Diagnosis-of-Diabetes-Mellitus-using-K-Nearest-Saxena-Khan/9112110d540f9d46cb6dcfaa160c9ac9c603382b

[22] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Proc. Comput. Sci.*, vol. 47, pp. 45–51, Jan. 2015, doi: 10.1016/j.procs.2015.03.182.

[23] S. R. Herrle, E. C. Corbett, M. J. Fagan, C. G. Moore, and D. M. Elnicki, "Bayes' theorem and the physical examination: Probability assessment and diagnostic decision making," *Academic Med.*, vol. 86, no. 5, pp. 618–627, May 2011, doi: 10.1097/ACM.0b013e318212eb00.

[24] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nursing*, vol. 43, no. 2, pp. 154–164, Apr. 2013, doi: 10.4040/jkan.2013.43.2.154.

[25] S. B. Wankhede, "Analytical study of neural network techniques: SOM, MLP and classifier—A survey," *IOSR J. Comput. Eng.*, vol. 16, no. 3, pp. 86–92, 2014, doi: 10.9790/0661-16378692.

[26] M. A. Sapon, K. Ismail, and S. Zainudin. (2011). *Prediction of Diabetes by Using Artificial Neural Network*. Accessed: Jun. 04, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Prediction-of-Diabetes-by-using-Artificial-Neural-Sapon-Ismail/c2c2696985801ba6b43b715f6aeaa4d272c438b4

[27] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Haryana, India: Morgan Kaufmann, 2011.

[28] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998, doi: 10.1109/5254.708428.

[29] G. Ridgeway, "The state of boosting," in *Computing Science and Statistics*, vol. 31. New York, NY, USA: Springer-Verlag, 1999, pp. 172–181.

[30] Y.-W. Kao, G.-H. Luo, H.-T. Lin, Y.-K. Huang, and S.-M. Yuan, "Physical access control based on QR code," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Oct. 2011, pp. 285–288, doi: 10.1109/CyberC.2011.55.

[31] P. C. Huang, Y. H. Li, C. C. Chang, and Y. Liu, "Efficient QR code authentication mechanism based on Sudoku," *Multimedia Tools Appl.*, vol. 78, pp. 26023–26045, Sep. 2019, doi: 10.1007/s11042-019-07795-8.

[32] K. Krombholz, P. Frühwirt, P. Kieseberg, I. Kapsalis, M. Huber, and E. Weippl, "QR code security: A survey of attacks and challenges for usable security," in *Human Aspects of Information Security, Privacy, and Trust* (Lecture Notes in Computer Science), vol. 8533, T. Tryfonas and I. Askoxylakis, Eds. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-07620-1_8.

[33] J. Zhang, D. Li, J. Jia, W. Sun, and G. Zhai, "Protection and hiding algorithm of QR code based on multi-channel visual masking," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4, doi: 10.1109/VCIP47243.2019.8966044.

[34] S. Tiwari, "An introduction to QR code technology," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Dec. 2016, pp. 39–44.

[35] M. Kayser, S. Brauer, G. Weiss, W. Schiefenhövel, P. Underhill, P. Shen, P. Oefner, M. Tommaseo-Ponzetta, and M. Stoneking, "Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea," *Amer. J. Hum. Genet.*, vol. 72, no. 2, pp. 281–302, Feb. 2003, doi: 10.1086/346065.

[36] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, Jul. 2018, doi: 10.3390/s18072183.

[37] D. T. Larose, *Data Mining Methods and Models*, 1st ed. Hoboken, NJ, USA: Wiley, 2006.

[38] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C.-H. Youn, "5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 16–23, Apr. 2018, doi: 10.1109/MCOM.2018.1700788.

RADWA MARZOUK received the B.S., M.S., and Ph.D. degrees in computer science from Cairo University, in 2000, 2005, and 2012, respectively. She is an Assistant Professor with the College of Computer and Information Sciences (CCIS), Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Since 2003, she has been a Faculty Member with the Department of Mathematics, Faculty of Science, Cairo University, Egypt. She has published several scientific papers in national, international conferences and journals. Her primary research interests include cryptography, computer security, coding theory, digital image processing, big data, bioinformatics, the IoT, AI, machine learning, computer networks, and applied mathematics.

ALA SALEH ALLUHAIDAN received the B.Sc. degree in computer science from Princess Nourah Bint Abdulrahman University Riyadh, Saudi Arabia, the M.Sc. degree in computer information systems from Grand Valley State University, MI, USA, and the Ph.D. degree in information systems and technology from Claremont Graduate University, CA, USA. She is currently an Assistant Professor with the Department of Information Systems, Princess Nourah Bint Abdulrahman University. Her research interests include health informatics, big data analytics, and machine learning.

SAHAR A. EL_RAHMAN (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Benha University, Cairo, Egypt, in 1997, 2003, and 2008, respectively. She is currently an Associate Professor with the Computer Systems Engineering Department, Faculty of Engineering-Shoubra, Benha University. Her research interests include artificial intelligence, machine learning, cryptography and information security, computer vision and image processing, human–computer interaction, big data, and cloud computing.