Received: 17 July 2021 | Revised: 29 August 2021 | Accepted: 9 September 2021
DOI: 10.37917/ijeee.17.2.14 Vol. 17| Issue 2| December 2021

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Towards for Designing Intelligent Health Care System Based on Machine Learning

Nada Ali Noori*, Ali A. Yassin
Department of Computer science, Education College for Pure Sciences, University of Basrah, Basrah, 61004, Iraq

Correspondence
*Nada Ali Noori
Department of Computer science
Education College for Pure Sciences,
University of Basrah, Basrah, Iraq
Email: pgs2190@uobasrah.edu.iq

Abstract

Health Information Technology (HIT) provides many opportunities for transforming and improving health care systems. HIT enhances the quality of health care delivery, reduces medical errors, increases patient safety, facilitates care coordination, monitors the updated data over time, improves clinical outcomes, and strengthens the interaction between patients and health care providers. Living in modern large cities has a significant negative impact on people's health, for instance, the increased risk of chronic diseases such as diabetes. According to the rising morbidity in the last decade, the number of patients with diabetes worldwide will exceed 642 million in 2040, meaning that one in every ten adults will be affected. All the previous research on diabetes mellitus indicates that early diagnoses can reduce death rates and overcome many problems. In this regard, machine learning (ML) techniques show promising results in using medical data to predict diabetes at an early stage to save people's lives. In this paper, we propose an intelligent health care system based on ML methods as a real-time monitoring system to detect diabetes mellitus and examine other health issues such as food and drug allergies of patients. The proposed system uses five machine learning methods: K-Nearest Neighbors, Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). The system selects the best classification method with high accuracy to optimize the diagnosis of patients with diabetes. The experimental results show that in the proposed system, the SVM classifier has the highest accuracy of 83%.

KEYWORDS: Diabetic prediction, Machine Learning, Electronic Health Record, Classification, Prediction Model

I. Introduction

The past decade has seen an explosion in digital information stored in Electronic Health Records (EHRs). During the same period, the machine learning (ML) community has witnessed large-scale developments in the field of EHRs and interest in patients' health care.

An EHR is an electronic version of a patient's medical history that the medical institution maintains over time and may include all the patient's clinical data care under a particular provider. The main contents of an EHR are demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory test data, and radiology reports [1].

The EHR is considered one of the advanced technologies that has changed the way the health care industry operates. Prior to the development of EHRs, medical records were 100% paper-based documents However, paper records hinder the health care environment due to limited accessibility, illegibility, inability to access files remotely, and the cost of storing huge files. EHRs can enhance patient care by improving the accuracy of medical records due to reducing the medical errors, providing information in a timely and safe manner to authorized users, preventing duplicate tests, reducing medication delays, and strengthening the relationship between patients and clinicians[2][3][4]

According to the World Health Organization, the most common chronic disease is diabetes [5]. Diabetes is a disease that occurs when blood sugar (glucose) levels are abnormally high. Blood glucose is the primary source of energy, which comes from the food we consume. Insulin, a hormone released by the pancreas, helps glucose absorption into the cells for energy. Sometimes the body does not generate enough—or any—insulin or does not use it properly. Glucose remains in the blood and does not reach the cells [6]. There are different types of diabetes, and each one requires special treatment. Not all types of diabetes are caused by overweight or following an unhealthy lifestyle. Some are present from childhood. Several factors cause diabetes, such as overweight, family history of diabetes, smoking, history of pressure, gestational diabetes, aging, and having a sedentary lifestyle[7]. Three common diabetes types can develop: Type I, Type II, and gestational diabetes [8].

Type I diabetes occurs when the body's insulin production is compromised. People with Type I diabetes who are insulin-dependent must take artificial insulin every day to stay alive. Type II diabetes, the most common type of diabetes, affects how the body uses insulin. While the body still produces insulin, unlike in Type I, the cells in the body do not respond to it as well as they once did [33]. According to the National Institute of Diabetes and Digestive and Kidney Diseases, it has strong links with obesity. Gestational diabetes occurs when a woman's body becomes less responsive to insulin during pregnancy. Gestational diabetes does not affect all women, and it typically goes away after the baby is born.

According to 2017 statistics, around 425 million people have diabetes. Every year, 2–5 million patients die due to diabetes, and by 2045 this will rise to 629 million [9].

Artificial intelligence (AI) is changing every aspect of our lives, including health care. The use of AI can significantly enhance the scope of diabetes care and make it more efficient. ML is considered one of the most important AI features that support the development of computer systems. With the current situation, there is an urgent need for ML to eliminate human efforts by supporting automation with minimum flaws [10].

The use of ML algorithms in health care is important for reducing the dangers due to the negative diagnosis and treatment. These algorithms work objectively to explain and synthesize the data in the EHR of each patient. The ML algorithms provide integration with clinical decision instruments, such as computerized alerts via SMS exported from mobile applications to clinicians and others who supply targeted and timely health information that improves clinical decisions. However, the ML algorithms can suffer from biases related to misclassification, missing data, and underestimating sample size, as well as measurement error. This potential biases that may be present in ML-based clinical decision support tools that use EHR data and suggests valuable solutions to issues of automation over-reliance, biased data-based algorithms, and algorithms that may have missed clinically meaningful information [11].

This paper describes the application of ML to clinical tasks based on EHR data in the health care system. We propose a variety of ML techniques and frameworks for diabetes, including information extraction, representation learning, outcome prediction, phenotyping, and identification. A brief description of the contribution of our proposed system is provided below.

1-We designed an intelligent medical system to analyze and predict the case of a diabetes patient based on a predictive model using ML algorithms such as K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR).

2-We created an EHR for every patient containing all the patient's medical information to facilitate the patient's diagnosis and quickly restore all their health information.

3-After entering the patient's laboratory test data, the system predicts whether the patient has diabetes or not.

4-We used the Pima Indian Diabetic Dataset (PIDD) with a total of 768 records for training and testing the ML algorithm, and then selecting the algorithm with the highest accuracy to predict the new cases.

5-The system prevents the doctor from giving the wrong prescription that conflicts with the patient's condition dependent on the patient's EHR, medicines that conflict with diabetes, and the foods to which they are allergic.

6-The system distributes the electronic clinic's management among the patient, the doctor, and the clinic's administrator while maintaining the confidentiality and privacy of the patient.

The remainder of this paper is organized as follows. Section II outlines the taxonomy of ML algorithms. Section III presents a literature review of the previous work on diabetes prediction. Section IV describes the proposed system. Section V presents the results of the experiment. Section VI presents the conclusions.

II. Background

ML algorithms consist of two main parts: classification and prediction. The first part is the process of recognizing, understanding, and grouping ideas and objects into preset categories or "sub-populations." Furthermore, several classification algorithms, such as KNN and NB, use the potential datasets to understand the problem and identify possible features and labels. These features represent the characteristics or attributes that affect the results of the label. The classification has two phases: learning and evaluation. In the first phase, the classifier trains its model on a given dataset, while in the second phase, the classifier’s performance is tested. The classifier's performance is evaluated based on various parameters, such as accuracy, precision, and recall.

The second part focuses on the prediction process, which uses the dataset's features to predict a new value based on the available data or used by any decision maker. There are many prediction algorithms, such as SVM [11][12]. Each of the ML algorithms used in our proposed system is described briefly below.

K-Nearest Neighbors Algorithm

KNN is a very popular, simple, and easy-to-understand ML algorithm. In many applications, KNN is used in numerous fields, such as image and video recognition, health care, political science, finance, and handwriting detection [13]. In KNN, K represents the number of nearest neighbors. The data is divided into groups or classes, and if new information needs to be classified, it finds the element's neighbors based on the majority of votes for the class label [14][34].

```
Algorithm K-Nearest Neighbor
 Input: the training dataset $D$, test item x, class label set $C$
 Output: the class $c_x$ of the test item $x$, $cx$ belongs to the $C$
   For each $y$ belongs to $D$ do
       calculate the Euclidean distance $D(y, x)$ between $y$ and $x$
   end for
   Select the subset $N$ from the data set $D$, the $N$ contains $k$ training samples,
   which are the $k$ nearest neighbors of the test item $x$
```

Naïve Bayes Algorithm

NB is a supervised learning algorithm for statistical classification and ideal for a large dataset. This technique uses the Bayes probability theorem to estimate unknown classes. The NB classifier has been successfully used in various applications, such as text classification, spam filtering, sentiment analysis, and recommendation systems [15].

NB's classifier assumes that the impact of a specific feature in a class is independent of other features. Even if these attributes are interdependent, these features are still separately considered. For a given class label, the prior probability is calculated. For each class, the probability for each attribute is determined, and after placing the values in the Bayes formula, the posterior probability is computed. The class that had the higher probability, given the input, belongs to the higher probability class [16]. Equation (1) simplifies the computation:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1)$$

- $P(h)$: The probability of a hypothesis.
- $P(D)$: The probability of the data.
- $P(h|D)$: The probability of the hypothesis given the data D.
- $P(D|h)$: The probability of the data given that the hypothesis was true.

```
Algorithm: Naive Bayes classifier
 Input: Training data $D$; Testing data $Y$;
Output: Estimated class $c$
Read training data $D$
Calculate posterior probabilities of training data in a class $p(d_i|c)$
Find the frequency of training data $Y$ for class $c$ $F_{ic}$
Find the number of occurrences $n$ of training data $Y_i$
   $W_i = f_i * n_i$
Find probability $p(c)$ of class $c$.
Get training data $Y$
if $(yj == Rj)$
   $prob(c|yj) = arg\ maxc\ p(c)\ \Pi\ p(yi|c)$
else
   $prob(c|yj) = arg\ maxc\ p(c)\ \Pi\ P(di|\ |c)wj$
Return  $c$ such that
$max\ \{prob(c|x1), prob(c|x2), \ldots, prob(c|xj)\ \}$.
End Algorithm Naive Bayes classifier
```

Logistic Regression Algorithm

LR is a common regression method used for solving binary classification problems. LR describes and estimates the relationship between one dependent binary variable and independent variables. It is a particular case of LR where the target variable is categorical. It changes the yield using the logistic sigmoid function to return a likelihood value [17].

```
Algorithm: Logistic Regression
  Input: Normalize (Dataset);
  Output: Estimated class $c$
  Repeat {
            $computedweight = gradient(parameters);$
            $Update\ computedweights;$
             $\}\ until\ convegence$
   $result =$
 $predictorvariables\ .update\ computedweights;$
   $Predict\_limit = sigmoid\ function(result);$
  Function gradient (predectattribute, targetattribute, weight)
            {calculate gradientdescent;
             Return weight+learningrate
```

Random Forest Algorithm

RF is a supervised learning algorithm used for classification and regression. RF has various applications, such as image classification, feature selection, recommendation engines, classifying loyal loan applicants, identifying fraudulent activity, and predicting diseases. It is based on the Boruta algorithm, which selects essential features in a dataset. RF is an ensemble method of decision trees generated on a randomly split dataset (also known as the forest). The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample[18] .

```
Algorithm: Random Forests

Input: Training data $D$; Testing data $Y$;
Output: Estimated class $c$
select targetattributes, and predictedattributes
Construct a decision tree for each data sample and get a prediction result from each decision tree.
Perform a vote for each predicted result.
Select the prediction result with the most votes as the final prediction result.
End Algorithm Random Forests
```

Support Vector Machine Algorithm

SVM has very high accuracy compared to other classification algorithms, such as LR. Its kernel trick handles nonlinear inputs, used in various applications such as face detection, intrusion detection, classification of emails, news articles, web pages, type of genes, and handwriting recognition.

The main objective is to find the best hyperplane that segregates the given dataset in the best possible way. The distance between the nearest points is known as the margin. The steps below can be followed for classification or regression tasks [19][20].

1-Preparation of the feature matrix: The feature matrix required for classification and regression is different. For classification, the data should be divided into two or more classes of extracted features with a label such as 1 or 0, indicating the class to which features belong. For regression, the input data and target data are often presented in separate columns of the matrix, in other words, training and testing data.

2-Selection of the kernel function: This is the most crucial step. The selection of suitable kernel functions depends on the nature of data and the type of application. The Gaussian kernel function used in the proposed system was the most successful kernel based on the results.

3-Parameter selection: Several parameters were selected to achieve the best performance, including the kernel functions parameters, the tradeoff parameter, and the insensitivity parameter. Selection of these parameters is not an easy task since there are no specific mathematical equations to provide a correct estimate of their values.

4-Training the algorithms: In the training step, the input and output data are defined. The nonzero values of the multipliers determine which of the input data will be the support vector. The margin of each class is determined by these vectors, which give the optimum hyperplane.

5-Classification/prediction: After determining the corresponding support vectors and multipliers, data can be appropriately classified. Any failure at this stage could be due to incorrect feature extraction, parameter estimation, or kernel selection. The best solution is to repeat the above steps to enhance the accuracy and reduce the error[21][22].

Evaluation Measurements

In this part, we evaluate the prediction results of each algorithm using variant evaluation metrics as described below.

1-Confusion Matrix: describes the performance of the model, and the output is a matrix, such as:

TABLE 1
Confusion matrix output

	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

2-Accuracy: the ratio of the total number of correct predictions to the total number of data samples.

3-Recall: the number of correct predicted results divided by the total number of relevant samples.

Precision: the ratio of accurate prediction results to the total number of positive expected results.

F-Score: the harmonic mean between recall and precision used to measure the accuracy of the test and tell us how robust and precise the classifier is. In other words, it tries to find the balance between recall and precision. All the measurement formulae are shown in Table (2).

TABLE 2
Measurement formulae

Measure	**Formula**
Accuracy (A)	$A = (TP+TN) / (TP + TN + FP + FN)$
Recall (R)	$R = TP / (TP + FN)$
Precision (P)	$P = TP / (TP + FP)$
F-Score (F)	$F = (2 \times P \times R) / (P + R)$
ROC	The area under the ROC curve

III. Literature Review

Many researchers have analyzed and predicted diabetes mellitus, and numerous prediction models have been developed and implemented using different ML algorithms.

Weifang Xu et al.[23] proposed a prediction model for Type II diabetes based on NB, Iterative Dichotomies 3 (ID3), AdaBoost, and RF. They studied some indicators (age, weight, waist, hip, …etc.) that may have an effect on diabetes. The results show that the RF algorithm has the highest prediction accuracy, and ID3 the lowest.

Messan et al. [24] explored the early prediction of diabetes using different data mining methods: Gaussian Mixture Model (GMM), Extreme Learning Machine (ELM), SVM, LR, and artificial neural network (ANN). They were applied to a diabetes dataset using only small data samples. The experiment results prove that Artificial Neural Network provides the highest accuracy than other techniques.

Perveen et al. [25] used Bagging, AdaBoost, and J48 (c4.5), and applied decision tree as a base learner with standalone data mining technique J48 to classify patients with diabetes. Their results indicate good accuracy and best performance.

Mounika et al. [26] used three classification algorithms: NB, OneR, and ZeroR. Factors such as smoking, diet maintenance, level of obesity, drug intake, and insulin deficiency were considered for predicting blood glucose levels among young and old patients. The Weks tools were implemented for estimating the accuracy and performance of each algorithm.

Mujumdar and Vaidehi [27] analyzed and compared the PIDD with a new dataset. Using KNN on only two attributes with high correlation, they applied LR, NB, RF, all used datasets and proposed a pipeline model to improve the accuracy of the classification model.

Sneha and Gangil [28]designed a prediction algorithm to identify the classifier that gives the closest clinical outcome results using ML algorithms such as DT, RF, and NB. They generalized only some features and deleted the features with low correlation. The decision tree and RF showed a good result compared to NB.

In [29]and[30], the researchers predicted the probability of diabetes in patients using KNN, NB, SVM, LR, and RF algorithms to determine the accuracy and misclassification rate. The estimation model obtained overcorrect and incorrect instances. The model results in [29] prove that LR had better accuracy and a lower misclassification rate, while the results in [30] show that SVM had the highest accuracy.

Meza-Palacios et al. [31]developed a fuzzy expert system (FES) for doctors to help them assess the kidney disease control in patients with Type 2 diabetes mellitus. Different features are used, such as blood glucose, serum creatinine, glomerular filtration rate (GFR), uric acid, age, dyslipidemia, and hypertension. Fuzzy systems help reduce the wrong treatment. Sabia et al. [32] proposed a medical decision support system (MDSS) focused on different diseases, including diabetes, consisting of a multi-layer classifier framework based on a combination of classifiers such as NB, RF, HiddenMarcoveVector(HMV), AdaBoost, KNN, LR, and SVM. The results show that the HMV algorithm provided the best accuracy. In this paper, we proposed a good intelligent system based on ML algorithms.

1- The system creates and manages the EHR for each patient and uses ML algorithms such as SVM to predict whether the patient has diabetes or not. If there is missing data, this will be compensated for by using the mean according to the diabetes database.

2-Our work has good results compared to related works in terms of accuracy, recall, precision, and ROC.

3-Our work aims to assist doctors to make better diagnoses and prevent medical errors by sending warning messages to doctors about the difficult cases of patient prescriptions.

4-Our work combines many techniques to develop an intelligent health care system based on ML algorithms to assist the patients to receive the best health care services and the doctors to use perfect diagnostic systems.

IV. THE PROPOSED SYSTEM

In this section, we discuss the main components of the proposed system represented by patients, doctors, and medical dataset management based on the ML algorithms. The proposed system consists of four main phases, which are outlined below (see Fig. 1).

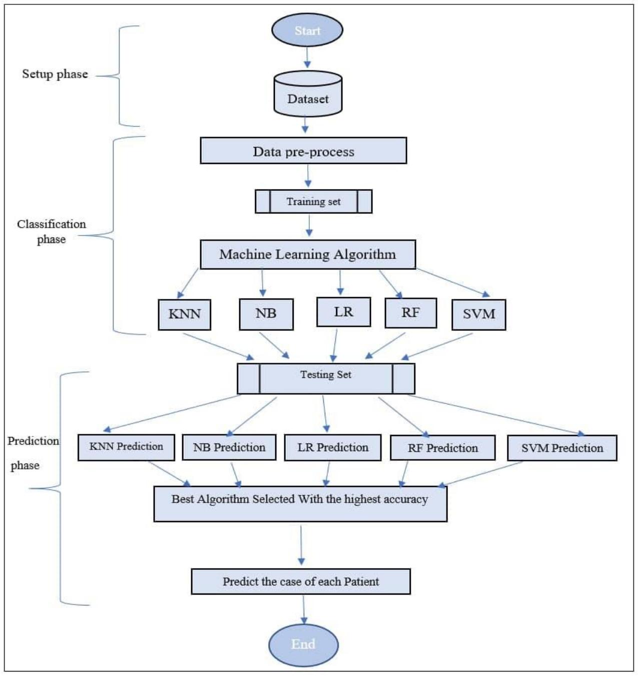

Fig. 1: System Work Flow.

1-Setup Phase

This phase is divides into two steps: data collection and data preprocessing.

Data collection

The dataset for the proposed system was obtained from the public UCI repository PIDD, which is open and freely available online. The dataset consists of 768 records and 8 attributes (insulin, pregnancies, blood pressure (BP), skin thickness, glucose, body mass index (BMI), diabetes pedigree function, and age), and one outcome class contains two values: 0 – nondiabetic, and 1 – diabetic [35].

Data preprocessing

This step handles the inconsistent dataset to obtain more precise and accurate results, such as deleting duplicate data and substituting missing values by computing the mean of all the attributes in the dataset since the medical data cannot have a value of zero and need to be normalized.

2-Classification Phase

The core phase of the proposed system includes building the model for diabetes prediction. We implemented different ML algorithms such as KNN, NB, LR, RF, and SVM. This phase involves two additional stages: the training stage and the evaluation stage.

Building the datasets

This stage involves dividing the data into three datasets: training set, validating set, and testing set. The training set is initially used to train the model and teach it how to process information. The validation set is used to finetune the model parameters and estimate the accuracy of the model. The testing set is used to evaluate the accuracy and performance of the model.

2.1 Training stage

The training set feeds the model to train the algorithm to learn to extract the relevant features and use them in the classification process. When the training stage is complete, the model is improved using the validation dataset. This involves altering or neglecting variables and tuning model settings until the model reaches a suitable accuracy level. The importance of this stage represents by extracting features based on the parameters suitable with each algorithm.

2.2 Evaluation stage

This stage aims to verify the proposed model by using accurate features and applying a testing set. Based on the feedback received, it may return to training the model to improve accuracy, adjust output settings, or deploy the model as needed. The following algorithm explains the main goals of the current phase. In addition, Fig. 2 illustrates the classification process workflow.

3- Prediction Phase

In this phase, we applied the ML algorithm with the highest accuracy to the new dataset to classify each patient into nondiabetic or diabetic. The result is 0 or 1 for each record.

The prediction phase works as follow:

- Defining a dataset that we use to train the data and extract the features.
- The fitting process reduces the prediction error by applying the training data to the machine learning algorithm and discover the mapping between the inputs and the output label.
- Prediction process: evaluate the model by using it to make predictions on the training dataset, then comparing the predictions to the expected label and calculate the classification report.
- Make a new prediction with new input and get one output by applying the new data of one row and multiple

columns to the ML algorithm, repeat the fitting and prediction process from the learning stage.

The new data row and the predicted class label (0 or 1) are applied to the database. Figure 3 demonstrates the prediction phase.

```
Algorithm: Model Building
 Input: Split dataset into training set and testing set
 Identify the algorithms that were used to create the model
 Classifier= [ KNN(), NB(), LR(), RF(), SVM()]
    For (index=1; index<=5; index++)
        Mode1=Classifier[index];
        Mode1.Fitting();
        Mode1.Predection();
        calculate (confusion matrix(index), accuracy(index),
recall(index), precession(index),
        F1_score(index), ROC (index));
        Print(confusion_matrix(index), accuracy(index),
precession(index),
        recall(index), F1_score(index), ROC (index))
 End Algorithm Model Building
```

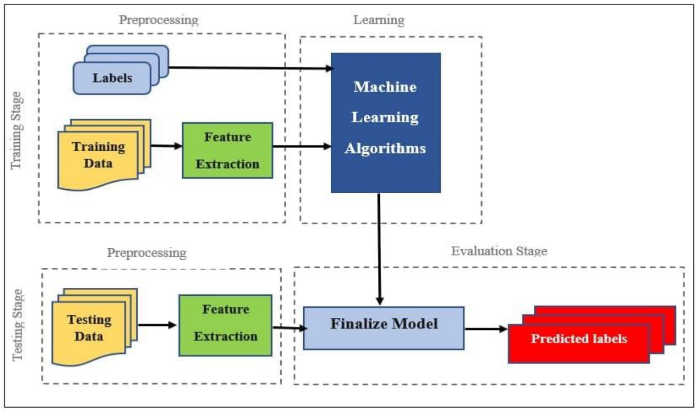

Fig. 2: Classification Phase Workflow

4- Health Care Phase

The patient wishes to check their diabetic case in the hospital. The following steps explain the mechanism of the current phase in our system.

Step 1: Add a new patient by creating their EHR, including personal information, laboratory test data, and information in relation to being allergic to food or environment.

Step 2: The prediction phase selects the best algorithm from the proposed system.

Step 3: The report is sent to the doctor for final treatment based on the following:

- Upon receiving the patient report, the doctor (d_i) writes the prescription.
- Our system checks whether the drug prescribed conflicts with the diabetic case and the food allergy.
- If it does, the system sends a warning message to the doctor to change the prescription.
- Otherwise, the prescription is fine and sent to the pharmacist.

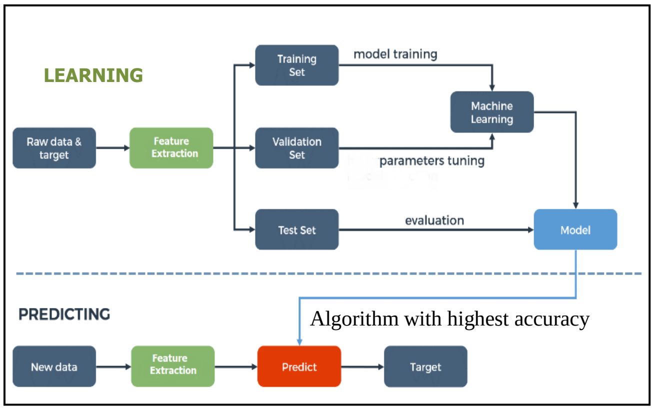

Fig. 3: Prediction phase workflow

Figure 4 demonstrate the health care scenario.

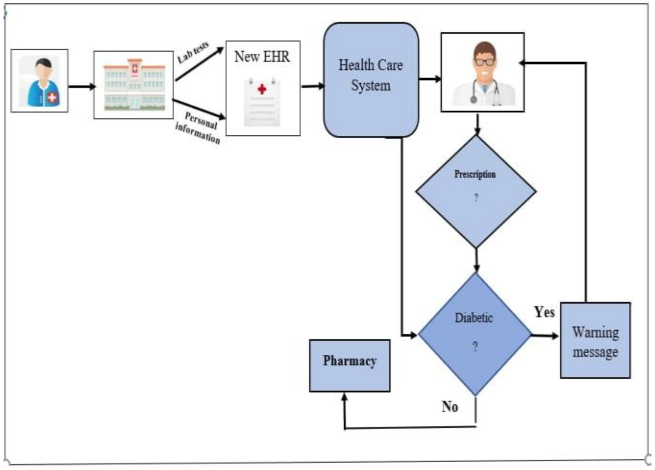

Fig. 4: Health Care Scenario.

V. Experiment Results

In this section, we explain the experimental results based on the major measurements outlined in section 2, as well as the comparison of different ML algorithms used in our work to evaluate the diagnosis of diabetes mellitus. In the proposed system, the model's performance is assessed based on correctly and incorrectly classified instances out of a total number of cases. The classifier performance is calculated based on evaluated metrics and can be calculated using Table (2). The performance evaluation of the classifiers is done through many evaluation measurements in term (%). The accuracy, confusion matrix, precision, recall, F-Score, and ROC of the different classifiers are compared in Table (3).

TABLE 3

Detailed measurements for K-nearest neighbor, Naive Bayes, logistic Regression, random forest, and SVM classifiers

Classification Algorithms	Confusion matrix	Accuracy (%)	F1 (%)	Precession (%)	Recall (%)	ROC (%)
KNN	84 12 21 37	79	69	76	64	76
NB	84 16 14 40	81	73	71	74	80
LR	140 15 30 46	81	67	75	61	75
RF	130 25 24 52	81	72	70	74	79
SVM	142 13 28 48	83	70	79	63	77

In Table (3), we observed that SVM achieved the highest accuracy among the classification algorithms with 82.4%. It was followed by RF with 81.4%, LR with 81.2%, NB with 81.1%, and KNN with 79.1%. Since we are dealing with medical data, we relied on accuracy as the main score for comparison, and the rest of the measurements are supportive of the work. Table (4) shows the comparison of the accuracy of the classification model in our work and previous researchers' work. Finally, the results showed that SVM is considered the best classification technique that met the requirements of our proposed system since it has the highest accuracy compared to the other classification techniques.

Each problem has a different analysis needing, depends on the type of the dataset, the selected parameters, and the classification problem. SVM gives the best results due to many reasons (1) when the dataset is small. (2) The problem is a binary classification. (3) Real valued features. (4) optimal margin gap between separating hyperplanes. (5) computationally more efficient because of using the Kernel trick in binary problems.

TABLE 4
Comparative analysis of model's accuracy (%)

Classifier	Mujumdar and vaidehi [25]	Sneha and Gangil [26]	Aishwarya Jakka Vakula rani [27]	Kishore et al. [28]	Proposed model
KNN	72.0	63.04	73.43	71.3	79.1
NB	67.0	73.48	75.52	-	81.1
LR	76.0	-	77.60	72.39	81.2
RF	72.0	75.39	74.30	74.4	81.4
SVM	68.0	77.73	65.63	73.43	82.4

Figures 5-9 explain the result of our system based on evaluation measurements

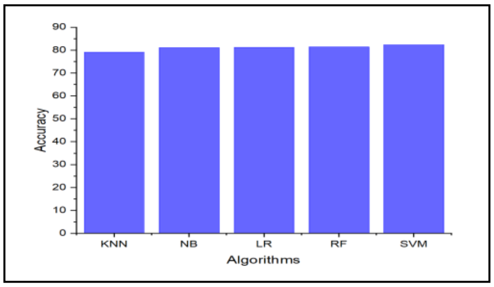

Fig. 5: Accuracy comparison of different algorithms

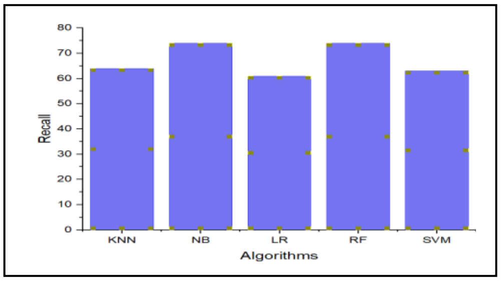

Fig. 6: Recall comparison of different algorithms

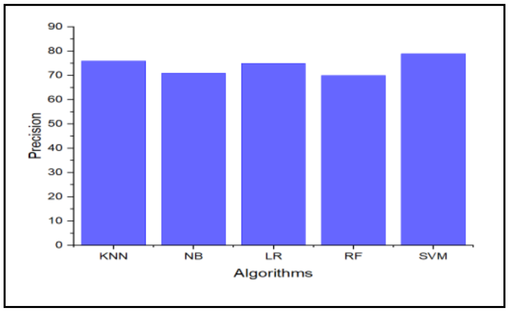

Fig. 7: Precision comparison of different algorithms

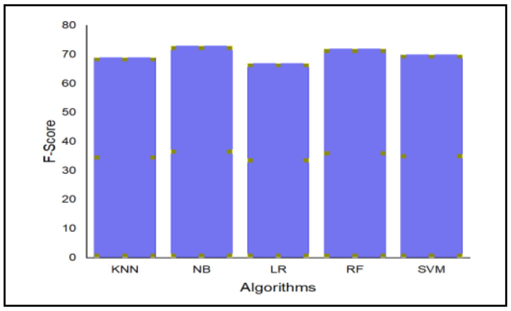

Fig. 8: F-score comparison of different algorithms

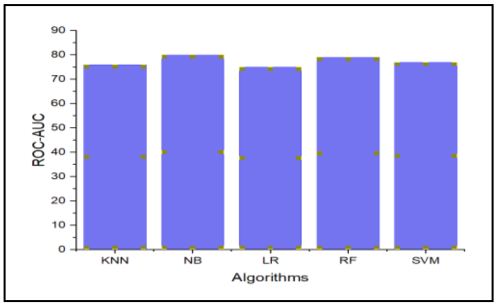

Fig. 9: ROC-AUC comparison of different algorithms

VI. Conclusion

The remarkable advances in information technology and the health care system have led to the production of a significant amount of data, such as medical information generated from huge EHRs. The Health Information Technology in medical clinics improves the quality of health care delivered by providing accurate patient records and allows doctors to understand the patient's medical history. In recent years, ML algorithms have shown significant results in detecting patterns associated with diseases and health conditions by studying thousands of health care records and other patient data. Diabetes is one of the real-world problems that need to be predicted and diagnosed early. This paper discusses a system designed to analyze and predict diabetes by comparing five ML algorithms assessed on different evaluation measurements and selecting the algorithm with the highest accuracy to predict the new case of the patient. The experimental results show that SVM has the highest performance among the classifiers used in the proposed system. The proposed system supports the health care sector by assisting doctors, patients, and medical enterprises. The system works as a monitor to prevent doctors from writing the wrong prescription and saves patients' lives.

Moreover, our work can substitute the missing data in the dataset using a suitable normalization method. Furthermore, our work does not depend on the fixed dataset, where each patient has an EHR, and their main information can be added to the dataset. We obtained good results compared to other related works. Additionally, our work can substitute the missing data in the dataset using a suitable normalization method. Additionally, our work doesn't depend on the fixed dataset, where each patient has EHR and can add his main information to the dataset. Thus, we gain good results compared with other related works. For Future work, we can use deep learning with big data to optimize the classification results. We can also add multiple data sets for different diseases like heart and kidney failure.

CONFLICT OF INTEREST

The authors have no conflict of relevant interest to this article.

REFERENCES

[1] G. S. Nelson, T. Technologies, and C. Hill, "A Practical Guide to Healthcare Data : Tips , traps and techniques," *Think. data*, vol. 1, no. August, pp. 1–20, 2017.

[2] C. H. Tsai, A. Eghdam, N. Davoody, G. Wright, S. Flowerday, and S. Koch, "Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content," *Life*, vol. 10, no. 12, pp. 1–27, 2020, doi: 10.3390/life10120327.

[3] L. Akhu-Zaheya, R. Al-Maaitah, and S. Bany Hani, "Quality of nursing documentation: Paper-based health records versus electronic-based health records," *J. Clin. Nurs.*, vol. 27, no. 3–4, pp. e578–e589, 2018, doi: 10.1111/jocn.14097.

[4] L. Waithera, J. Muhia, and R. Songole, "Impact of Electronic Medical Records on Healthcare Delivery in Kisii Teaching and Referral Hospital," *Med. Clin. Rev.*, vol. 03, no. 04, pp. 1–7, 2017, doi: 10.21767/2471-299x.1000062.

[5] D. Meetoo, "Chronic diseases: the silent global epidemic.," *Br. J. Nurs.*, vol. 17, no. 21, pp. 1320–1325, 2008, doi: 10.12968/bjon.2008.17.21.31731.

[6] M. Güemes, S. A. Rahman, and K. Hussain, "What is a normal blood glucose?," *Arch. Dis. Child.*, vol. 101, no. 6, pp. 569–574, 2016, doi: 10.1136/archdischild-2015-308336.

[7] A. Khan and S. Khan, "Causes, Complications and Management of Diabetes Mellitus," no. August, 2017.

[8] R. Goldenberg and Z. Punthakee, "Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome," *Can. J. Diabetes*, vol. 37, no. SUPPL.1, pp. 8–11, 2013, doi: 10.1016/j.jcjd.2013.01.011.

[9] M. Abusaib *et al.*, "Iraqi Experts Consensus on the Management of Type 2 Diabetes/Prediabetes in Adults," *Clin. Med. Insights Endocrinol. Diabetes*, vol. 13, 2020, doi: 10.1177/1179551420942232.

[10] S. Ellahham, "Artificial Intelligence: The Future for Diabetes Care," *Am. J. Med.*, vol. 133, no. 8, pp. 895–900, 2020, doi: 10.1016/j.amjmed.2020.03.033.

[11] M. A. Jabbar, S. Samreen, and R. Aluvalu, "The future of health care: Machine learning," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 23–25, 2018, doi: 10.14419/ijet.v7i4.6.20226.

[12] P. Pundir, V. Gomanse, and N. Krishnamacharya, "Classification and Prediction techniques using Machine Learning for Anomaly Detection .," *Pdfs.Semanticscholar.Org*, vol. 1, no. 4, pp. 1716–1722, 2011, [Online]. Available: https://pdfs.semanticscholar.org/267d/0ba8de46c022bf9ffd6af4cd0c4b403798ea.pdf.

[13] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.

[14] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, 2016, doi: 10.21037/atm.2016.03.37.

[15] J. Kazmierska and J. Malicki, "Application of the Naïve Bayesian Classifier to optimize treatment decisions," *Radiother. Oncol.*, vol. 86, no. 2, pp. 211–216, 2008, doi: 10.1016/j.radonc.2007.10.019.

[16] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.

[17] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergol. Immunopathol. (Madr).*, vol. 39, no. 5, pp. 295–305, 2011, doi: 10.1016/j.aller.2011.05.002.

[18] J. You, S. A. S. van der Klein, E. Lou, and M. J. Zuidhof, "Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision feeding system," *Comput. Electron. Agric.*, vol. 175, no. June, p. 105526, 2020, doi: 10.1016/j.compag.2020.105526.

[19] Y. J. Ccoicca, "Applications of Support Vector Machines in the Exploratory Phase of Petroleum and Natural Gas: a Survey," *Int. J. Eng. Technol.*, vol. 2, no. 2, p. 113, 2013, doi: 10.14419/ijet.v2i2.834.

[20] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, 2010, doi: 10.1109/TITB.2009.2039485.

[21] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 2, pp. 728–734, 2015, doi: 10.1109/JBHI.2014.2325615.

[22] R. G. Brereton and G. R. Lloyd, "Support Vector Machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010, doi: 10.1039/b918972f.

[23] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," *Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, AEEICB 2017*, pp. 382–386, 2017, doi: 10.1109/AEEICB.2017.7972337.

[24] M. Komi, J. Li, Y. Zhai, and Z. Xianguo, “Application of data mining methods in diabetes prediction,” *2017 2nd Int. Conf. Image, Vis. Comput. ICIVC 2017*, no. S Ix, pp. 1006–1010, 2017, doi: 10.1109/ICIVC.2017.7984706.

[25] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[26] M. Mounika, S. Suganya, B. Vijayashanthi, and S. Anand, “Predictive analysis of diabetic treatment using classification algorithm,” *Ijcsit*, vol. 6, no. 3, pp. 2502–2505, 2015, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.734.9118&rep=rep1&type=pdf.

[27] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[28] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.

[29] R. Deo and S. Panigrahi, “Performance Assessment of Machine Learning Based Models for Diabetes Prediction,” *2019 IEEE Healthc. Innov. Point Care Technol. HI-POCT 2019*, no. 11, pp. 147–150, 2019, doi: 10.1109/HI-POCT45284.2019.8962811.

[30] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” *Procedia Comput. Sci.*, vol. 167, no. 01, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[31] R. Meza-Palacios, A. A. Aguilar-Lasserre, E. L. Ureña-Bogarín, C. F. Vázquez-Rodríguez, R. Posada-Gómez, and A. Trujillo-Mata, “Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus,” *Expert Syst. Appl.*, vol. 72, no. February 2019, pp. 335–343, 2017, doi: 10.1016/j.eswa.2016.10.053.

[32] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, “HMV: A medical decision support framework using multi-layer classifiers for disease prediction,” *J. Comput. Sci.*, vol. 13, pp. 10–25, 2016, doi: 10.1016/j.jocs.2016.01.001.

[33] https://www.who.int/health-topics/diabetes

[34] M. Z. Al-Faiz, A. A. Ali, and A. H. Miry, “A k-nearest neighbor based algorithm for human arm movements recognition using EMG signals,” Iraqi Journal for Electrical and Electronic Engineering, vol. 6, no. 2, pp. 159–167, 2010, doi: 10.37917/ijeee.6.2.12.

[35] https://www.kaggle.com/uciml/pima-indians-diabetes-database