Healthcare Data Analytics and Clinical Decision Support System, Special Case of Diabetes Mellitus

Divya Patel[1], Niteen Patel[2]
[1]*Student, Department of Electronics and Communication*
[2]*Associate Professor, Department of Electronics and Communication*
[12]*Sarvajanik College of Engineering & Technology, Surat, India*
(E-mail: [1]divyapatel8347@gmail.com, [2]niteen.patel@scet.ac.in)

Abstract—Diabetes Mellitus is found to be the fourth leading cause of global death by disease. Type 2 Diabetes Mellitus (T2DM) is the most common form of diabetes — around 90% of people with diabetes have type 2 diabetes. Early detection, diagnosis, and cost-effective treatments can save lives and prevent or significantly delay destructive diabetes-related complications. In this paper, a prediction model is proposed based on data analytics techniques for predicting T2DM, which comprised of two parts, First, the Clustering Algorithms like Improved K-means and Fuzzy C-means. Second, the Classification Algorithms like Logistic Regression and Gradient Descent.

The Pima Indians Diabetes Dataset was utilized to compare the results from other researchers. The result shows that the proposed model has reached better accuracy compared to other previous studies that mentioned in the literature. On the basis of the result, it can be proven that the proposed model would be helpful in Type 2 diabetes diagnosis.

Keywords— *Data Analytics, Prediction Model, Type2 Diabetes, Clustering Algorithms, Classification Algorithms*

I. INTRODUCTION

Diabetes Mellitus (DM) is a group of metabolic diseases that affect the body's ability to regulate blood glucose levels. It is characterized by elevated levels of blood glucose, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves. Primarily, there are three types of diabetes which hold Type-1 DM (T1DM), Type-2 DM (T2DM) and Gestational 0diabetes.T1DM usually diagnosed in children and young adults in which the body does not produce enough insulin. Only 5% of diabetes people have T1DM. T2DM is the most common form of diabetes; about 90% of diabetes people have T2DM. In T2DM, Blood glucose levels to rise higher than normal. It is also called hyperglycemia in which the body produces insulin but can't use it well. Gestational diabetes is a temporary condition that occurs during pregnancy [1].

According to the eighth edition of IDF (International Diabetes Federation) Diabetes Atlas, approximately 425 million adults (20-79 years) in 2017 were living with diabetes, with dramatic increases seen in countries all over the world and due to that 4 million people were deaths. By 2045, it is estimated that 629 million people will suffer from diabetes. T2DM constitutes the majority of all diabetes. The numbers of people at risk of developing T2DM were 352 million [2]. As a consequence, T2DM is a severe health issue for the whole world. If we could Diagnose and prevent diabetes as early as possible, millions of lives might be saved.

In order to research the high-risk group of DM, we need to utilize advanced information technology. Therefore, Data Analytics an appropriate study field for us. Data Analytics refers to the analysis of unstructured data to extract the knowledge hidden in every single bit of data for making better decisions and predictions [3]. Analytics solutions are of four types: (a) Descriptive, which uses business intelligence and data mining to ask: "What has happened?" (b) Predictive, which uses statistical models and forecasts to ask: "What could happen?" (c) Prescriptive, which uses optimization and simulation to ask: "What should we do?" (d) Diagnostic, examines data to answer "Why did it happen?" [4]. In Healthcare, these four types of analytics were used in different applications area i.e., clinical decision support, healthcare administration, privacy and fraud detection, mental health, public health, and pharmacovigilance.

The main goal of the study is to take the historical data of the patient and improve the patient current outcomes. Predictive analytics is a powerful tool in this regard. Predictive Analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics applies many techniques from data mining, statistics, modelling, machine learning, and artificial intelligence to investigate current findings to make predictions about the future. Predictive Analytics is supporting different segments of health care life sciences and providers. It aims in diagnosing the diseases accurately, enhancement of patient care, resource optimization and also improves clinical outcomes. Predictive Analytics helps organizations to prepare for the health care by optimizing the cost. The accomplishment of predictive analytics in this industry is likely to provide proficient outcome by improving the service quality. Predictive Analytics have the future to transform the health care industry [5] [6].

Among application areas, the clinical decision support system (CDSS) had the highest application of predictive analytics as many studies in this area are involved in risk and morbidity prediction of diabetes, chest pain, heart attack, and other diseases [7]. The purpose of a CDSS is to assist healthcare providers, enabling an analysis of patient data and using that information to aid in formulating a diagnosis. A CDSS offers information to clinicians and primary care providers to improve the quality of the care their patients receive [7].

The high prevalence of DM, and the rapidly growing number of patients with DM, along with the rising costs of care, the predictable number of deaths and medical errors, poses the need to move from a reactive to a preventive approach in diabetes care and to shift the importance from the disease to wellness. Therefore, it is essential to develop an expert predictive healthcare decision support system using data analytics techniques that can classify patients into either suspected patients or confirmed patients from the first examination time for the high-risk DM group.

Section II presents the related work of prediction based diagnosis in the group of diabetes. Section III details the prediction model, dataset and algorithms. Section IV describes the results of the experiment and comparative analysis. Section V concludes the paper with some directions for future work.

II. RELATED WORKS

The extensive availability of healthcare data in the past decade, as well as advancements in the area of data mining and machine learning, has generated an interesting field of healthcare analytics. Many algorithms and toolkits have been created and studied by researchers. These have highlighted the tremendous potential of this research field. The development of CDSSs by data analysts with the aid of clinical experts' knowledge has eased the burden on physicians and clinicians and smoothed clinical procedures. Analyzing healthcare data and applying machine learning techniques in this area have several benefits: patients can be stratified based on the severity of a particular disease or condition and, consequently, suitable treatments can be provided for each group; risk factors of different diseases can be identified, leading potentially to better health management; and diseases can be detected at early stages, allowing for appropriate interventions and treatments [8]. In this section, a few significant works that are intimately related to the proposed issue are presented.

Based on several studies, we found that a commonly used dataset was the Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database [9]. Patil [10] and Chen [11] proposed a hybrid prediction model (HPM), which used a simple K-means clustering algorithm aimed at removing incorrectly classified instances, i.e. pattern extracted from original data and used the J48 decision tree as a classifier for classification, with 92.38% [10] and 90.04% [11] classification accuracy using data mining. Both these prediction models [10] [11] used deleting overmuch data from the original dataset which cause inaccurate experiment results. So the original dataset samples were reduced than the original one due to deleting noisy data.

In order to obtain more useful and meaningful data from the raw dataset, we realized that the preprocessing methods and parameters should be chosen rationally. Also, improvements in Machine learning algorithms will give rise to prediction accuracy. Han Wu [12] proposed a novel model based on data mining techniques which were comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. Because of using improvement in the k-means algorithm and using series of preprocessing procedures, Han Wu [12] obtain 95.42% classification accuracy, which is 3.04% higher accuracy of prediction than those of other researchers. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material and the Waikato Environment for Knowledge Analysis (WEKA) as a toolkit.

However, the prediction accuracy and data validity were not high enough for a realistic application. So in this study, we try to improve the logistic regression algorithm by the process of minimizing a function by following the gradients of the cost function (Gradient Descent). We used MATLAB as a toolkit, the programming language developed by MathWorks which is a suitable platform for predictive analysis as it is rich in Machine learning libraries and is easy to implement new features. MATLAB is not very popular when it comes to data science but it is one of the languages that many people consider for learning data science very easily [13].

III. PROPOSED WORK

In this section dataset description, data pre-processing, clustering and classification algorithms are discussed. The clustering algorithms like improved K-means and fuzzy C-means (FCM) and the classification algorithms like logistic regression and gradient descent are taken for analysis. All the experimental processes have been completed using the MATLAB. The proposed model is shown in Fig. 1.

1. Dataset Description

The Pima Indian Diabetes (**PID**) Dataset consists of information of 768 patients in which there are 268 tested_positive instances and 500 tested_negative instances, coming from a population near Phoenix, Arizona, USA. Tested_positive and tested_negative indicate whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. These data contain personal health data as well as results from medical examinations.

Figure 1: Block Diagram of the Proposed Method

The detailed attributes in the dataset are listed as follows,

- Number of times Pregnant(preg)
- Plasma Glucose concentration a 2 hours in an oral glucose tolerance test(plas)
- Diastolic Blood Pressure (mm Hg)(pres)
- Triceps Skin fold Thickness (mm)(skin)
- 2-Hour serum Insulin (mu U/ml)(insu)
- Body Mass Index (weight in kg/(height in m)^2)(bmi)
- Diabetes Pedigree Function(pedi)
- Age: Age (years)(age)
- Outcome: Class variable (0 or 1)(class)

2. *Data Pre-processing*

Data pre-processing were applied in order to improve the quality of the prediction and the efficiency of the predictive analytics process.

a) We determined that the number of pregnancies has little connection with DM. Therefore, transformed this numeric attribute into a nominal attribute. Value of Pregnancies in terms of 1 and 0. i.e., Pregnant=1, non-pregnant=0
The complexity of the dataset was reduced by this process.

b) There are some missing and incorrect values in the dataset due to errors or deregulation. Most of the inaccurate experimental results were caused by these meaningless values. For example, in the original dataset, the values of diastolic blood pressure and BMI could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values,
Used the means from the training data to replace all missing values

c) After the above steps were applied, the unsupervised normalize filter for attribute was used to normalize all the data by using (1):

$$\text{Value}' = \frac{\text{value} - X'}{s} \quad (1)$$

Where, x' = Mean or average value for the variable
s = Standard deviation for the variable
Value' = New normalized value

This avoids the complexity of calculation and accelerates the speed of the operation.

3. *Model Description*

The predictive model consists of a double level algorithm. In the first level, we used the clustering algorithm to remove incorrectly clustered data and optimized data was used as input to classification algorithms.

3.1. *Improved K-means Clustering Algorithm*

In this study, an improved K-means algorithm is used to remove incorrectly clustered data. First, we see the procedures of the simple K-means cluster algorithm are as follows [14]:

a) Show all objects. Select K from provided N as the number of the initial cluster center. According to the value of K, partition the dataset into K subsets. Then for this each subset calculates the mean, which is treated as the initial cluster Centre.

x = (x1, x2, …, xn) = set of objects (Dataset)

S = {S1, S2, …, Sk} = partition of n observations into k (≤ n) sets
mi = Mean of points in Si

Given an initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

b) Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

Calculate the distance between each object and cluster center. Cluster every object to the nearest cluster according to the distance using (2):

$$S_i^{(t)} = x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \quad (2)$$

$$\forall j, 1 \le j \le k$$

Where each x_p is assigned to exactly one $s^{(t)}$, even if it could be assigned to two or more of them.

c) Update step: Recalculate every cluster center to verify whether they are changed using (3):

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} Xj \quad (3)$$

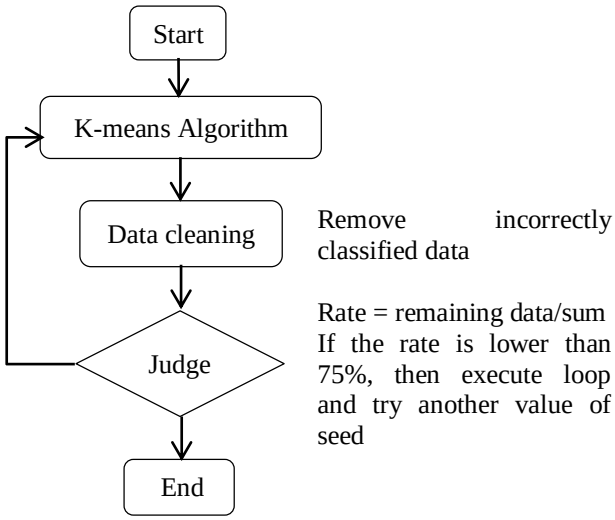

Figure 2: Improved K-means Algorithm

d) Circulate step (b) and step (c) until the new cluster center is the same as the original one, i.e., convergence and end of the algorithm

Stopping criterion:

a) When the maximum number of iterations has been reached
b) When improvement in results is less than a threshold

Advantages of simple K-means: Rapid convergence

Disadvantages of simple K-means: Sensitivity to initialization: an objective function of K-means is not convex, so it may have numerous local minima. i.e., the performance of this algorithm is dependent on the initial choice of cluster centers. Euclidean distance is sensitive to noise and outlier data. So, an improved k-means algorithm is used.

As shown in Fig 2, the incorrectly classified data obtained after applying the k-means algorithm were removed in the improved k-means algorithm. Then calculated the rate [12] using the formula expressed as (4),

$$Rate = \frac{Remaining\ data}{Sum} \tag{4}$$

If the rate was higher than 75%, then we moved to the next level. Otherwise, it should exit the loop and try another seed value.

Table 1: Result of the improved K-means Algorithm

No	Label	Count
1	Diabetic	295
2	Non-Diabetic	473

Table1 shows the results of the improved 2-means clustering in 2 clusters which is Diabetic and Non-Diabetic means positive class and negative class respectively. After the removal procedure, we obtained 589 correctly classified patients using the improved k-means algorithm, which all served as input to the Classification algorithm.

3.2. Fuzzy C-means Clustering Algorithm

To determined data set X = {x1, x2, ... xn} ⊂ Rs the FCM is a iterative process that divides X in c clusters. The result of clustering is expressed by the degrees membership in the matrix μ, where μ_{ij} is the membership degree of the object x_i the j-th cluster [14]. The algorithm FCM attempts to find a partition fuzzy that represents the data structure, minimizing the objective function defined by (5)

$$J = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}^{m}\ d(x_i;\ c_j)^2 \tag{5}$$

And with the restrictions by (6)

$$\sum_{j=1}^{c} \mu_{ij} = 1\ ,$$
$$\forall_i \in \{1, \ldots, n\}; 0 < \sum_{i=1}^{n} \mu_{ij} < n, \forall_j \in \{1, \ldots, c\} \tag{6}$$

Where, n = number of data
c = number of clusters
d (x_i;c_j) = distance between x_i and c_j
m = [1.25; 2]; m>1 is the fuzziness value that controls how much fuzzy is the partition.
$x_i \in R^s$(i =1,...,n) = vector of data, in which each position in the vector represents an attribute;
$c_j \in R^s$(j =1,...,c) = the centroid cluster of the j-th cluster

The main steps of the algorithm FCM are described below:

Step 1: Initialize the matrix partition μ with continuous random numbers at the interval [0, 1];

Step 2: Calculate the centroid of the cluster j using (7),

$$c_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m}\, x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \tag{7}$$

Table 2: Result of fuzzy c-means

No	Label	Count
1	Diabetic	358
2	Non-Diabetic	410

Step 3: Calculate an initial value to J as given by (5)

Step 4: Calculate the membership degrees on matrix fuzzy μ of the following form:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d(x_i\,;\,c_j)}{d(x_i\,;\,c_k)}\right)^{\frac{2}{m-1}}} \tag{8}$$

Step 5: If d (J_U; J_A) ≤ ε Stop. If not, go back to step 2.

The criterion of convergence is the threshold ε>0. Other criterion of possible halt is when a number of prefixed iterations are executed.

Table 2 shows the clustering results of fuzzy C-means in which Diabetic and Non-Diabetic means positive class and negative class respectively. We obtained 558 correctly classified patients using the fuzzy c-means algorithm, which are all applied as input to the classification algorithm.

3.3. Logistic Regression Classification Algorithm

Logistic regression is a discriminative, linear model for binary classification [15]. The logistic regression algorithm is based on the linear regression model expressed as (9)

$$\begin{aligned} Y &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \\ &= [\beta_1, \beta_2, \ldots \beta_m]^T [X_1, X_2, \ldots X_m] = \boldsymbol{\beta}^T X \end{aligned} \tag{9}$$

That is, it models the probability distribution Pr(Y | X) where Y is the class label of the item (either -1 or 1), and X is its feature representation. Once Pr(Y | X) is learned, the model will classify a new item as (10),

$$\begin{aligned} &\Pr(Y = +1|X) \approx g(\boldsymbol{\beta}^{T} X) \quad \text{and} \\ &\Pr(Y = -1|X) = 1 - \Pr(Y = +1|X) \end{aligned} \tag{10}$$

Where the probability distribution Pr(Y | X) is represented as (11)

$$\Pr(Y = +1|X) = \frac{1}{1 + e^{-\beta^T x}} = \sigma(\boldsymbol{\beta}^T X) \tag{11}$$

Where σ = the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$

β = the weight-vector.
Y = the patient is diabetic or Not
X = Independent variables
= Represent the 8 attributes in the original dataset

3.4. Gradient Descent Classification Algorithm

In Logistic Regression, given a training set, how do we pick, or learn, the parameters β? One reasonable method seems to be to make h(x) close to y. To formalize this, we will define a function that measures, for each value of the β's, how close the $h(x^{(i)})$'s are to the corresponding $y^{(i)}$'s [16]. We define the cost function as (12):

$$J(\beta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log\left(h_\beta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\beta\left(x^{(i)}\right)\right)\right] \tag{12}$$

We want to choose β so as to minimize J(β). Specifically, let's consider the gradient descent algorithm, which starts with some initial β, and repeatedly performs the update using (13):

$$\beta_j := \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\beta) \tag{13}$$

Here, α is called the learning rate. This is a very natural algorithm that repeatedly takes a step in the direction of steepest decrease of J. In order to implement this algorithm, we have to work out what is the partial derivative term on the right hand side. For a single training example, this gives the update rule as (14):

$$\beta_j := \beta_j - \alpha\left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right) x_j i \tag{14}$$

Repeat until convergence {

$$\beta_j = \beta_j - \alpha \sum_{i=1}^{m} \left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right) x_j i \tag{15}$$

}

IV. EXPERIMENTAL RESULT

All the experiments are performed using the Statistics and Machine Learning Toolbox and the Fuzzy Logic Toolbox in MATLAB.

The experiments were performed using four different prediction models which are developed using (1) Improved K-means as clustering and logistic regression as classification algorithm (2) Improved K-means as clustering and gradient descent as classification algorithm (3) Fuzzy C-means as clustering and Logistic regression as classification algorithm (4) Fuzzy C-means as clustering and gradient descent as classification algorithm. Here, Gradient descent is used as the optimization of the cost function used by the logistic regression algorithm.

Even if the fuzzy C-means algorithm shows improve accuracy compare to other researchers' work, we go with the choice of improved K-means algorithm, because the number of incorrectly clustered instances is high in the fuzzy C-means algorithm. So, for comparative analysis with other researchers work we proposed a model using improved k-means and the

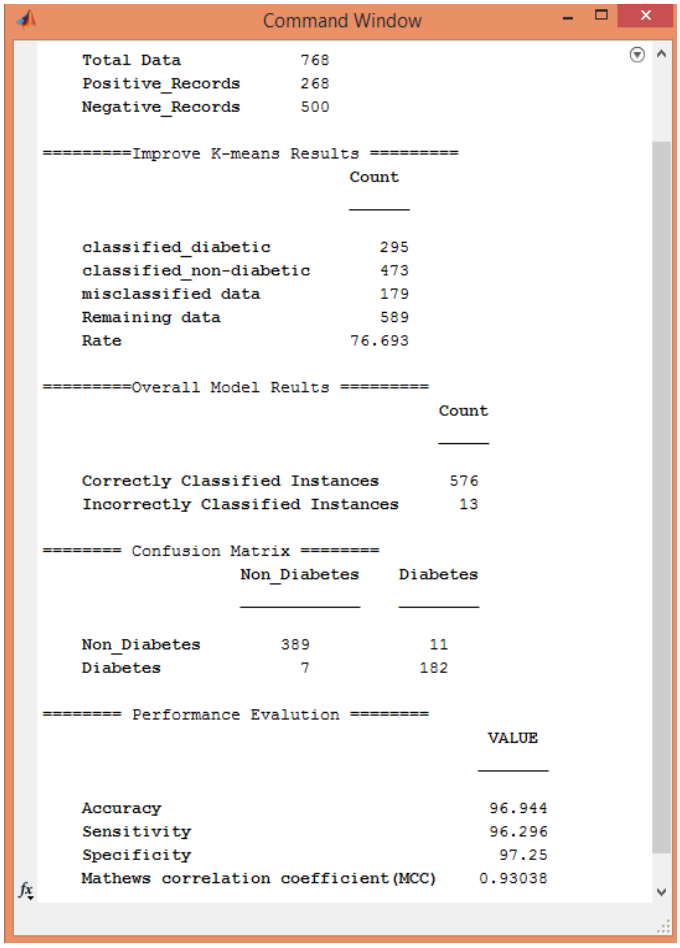

Figure 3: The result of the Experiment.

Gradient descent algorithm. The experiment result of this model is as shown in Fig.3.

We analyzed and evaluated our model based on the following aspects.

A. *K-fold Cross Validation*

In order to have a good measure performance of the classifier, k-fold cross-validation method has been used. For a k-fold cross-validation, each single example occurs exactly k-1 times as a training example. Hence, the time needed to compute the statistics of all test examples is reduced by a factor k-1 compared to running the original algorithm k times. We used 10 fold cross-validations in our Prediction Model. It can reduce the bias associated with random sampling method. This can speed-up the algorithm and there are no changes in the computational complexity of the algorithm.

B. *Accuracy, Specificity and Sensitivity*

Generally, the process of prediction has the four different possible outcomes shown in Table 3.

a) True positives (TP): It occurs when the outcome is correctly predicted as same as the actual one.
b) True negatives (TN): It occurs when the outcome is correctly predicted as same as the actual one.
c) False positives (FP): It occurs when the outcome is incorrectly predicted as positive when it is actually negative.
d) False negatives (FN): It occurs when the outcome is incorrectly predicted as negative when it is actually positive.

In this study, the following equations (16), (17) and (18) are used to measure the accuracy, sensitivity and Specificity respectively.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (17)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (18)$$

The Mathews correlation coefficient (MCC) is used as a measure of the quality of binary classifications is given by (19)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (19)$$

C. *Confusion Matrix*

The confusion matrix displays four different outcomes (TP, TN, FP, and FN) of this study as shown in Table 4. Column A presents the tested positive results, and Column B presents the tested negative results. The first row shows the predicted results for the positive class, and the second row shows the predicted results for the negative class. For a classifier that has good accuracy, ideally most of the outcomes would be represented along the diagonal of the confusion matrix with the rest of the entries being zero or close to zero.

Table 3: Confusion Matrix

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	TP	FN
	No	FP	TN

Table 4: Confusion Matrix of the experiment

A	B	Classified
389	11	A=tested_negative
7	182	B=tested_positive

D. Comparative Analysis

In Table 5, it is sawing comparisons of this study with other Researchers. From this we can say that proposed model with Improve K-means and gradient descent algorithms display high accuracy than model proposed by other Researchers.

Table 5: Comparative Analysis

Reference	Method	Accuracy (%)
Patil [10]	K-means clustering + J48 decision tree algorithm	92.38
Chen [11]	K-means clustering + J48 decision tree algorithm	90.04
Han Wu [12]	Improve K-means + Logistic regression	95.42
This Study	Fuzzy c-means + Logistic regression	96.237
	Fuzzy c-means + Gradient Descent	98.208
	Improve K-means + Logistic regression	95.416
	Improve K-means + Gradient Descent	96.944

V. CONCLUSION

The main problem we solved is improving the accuracy of the prediction model. In this paper, we conclude that our proposed model exhibiting higher prediction accuracy than other researchers' experimental results. And the Gradient Descent algorithm we proposed contributed a lot to the prediction model. With the rapidly growing demand for medical data analytics, the proposed model can be fairly helpful to the researchers and doctors for their decision-making on the patients as by using such an efficient model they can make more accurate decisions.

For future work, it is necessary to bring in the hospital's real and latest patients' data for continuous training and optimization of our proposed model. The quantity of the dataset should be large enough for training and predicting. Some advanced algorithms and models should be applied in the research of DM. This will help to limit the growth rate of diabetes and eventually decrease the risk of developing DM.

REFERENCES

[1] 'About Diabetes'. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes.html

[2] International Diabetes Federation. *IDF Diabetes Atlas*, *8th edn*. Brussels. Belgium: International Diabetes Federation, 2017, [Online]. Available: http://www.diabetesatlas.org

[3] Ahmed M, Pathan AS. Data Analytics: Concepts, Techniques, and Applications. CRC Press; 2018 Sep 21.

[4] 'Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics', [Online]. Available: https://www.ibm.com/downloads/cas/3V9AA9Y5

[5] 'Predictive Analytics', [Online]. Available: https://www.ibm.com/in-en/analytics/predictive-analytics

[6] Islam MS, Hasan MM, Wang X, Germack HD. A systematic review on healthcare analytics: Application and theoretical perspective of data mining. InHealthcare 2018 Jun (Vol. 6, No. 2, p. 54). Multidisciplinary Digital Publishing Institute.

[7] 'Clinical decision support system (CDSS)', [Online]. Available: https://searchhealthit.techtarget.com/definition/clinical-decision-support-system-CDSS

[8] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Information Science and Systems 2 (2014) 1.

[9] 'Pima Indians Diabetes Database', [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[10] B. M. Patil; R. C. Joshi; Durga Toshniwal.,"Hybrid prediction model for Type-2 diabetic patients" in Expert Systems with Applications, Vol: 37, Issue: 12, Page: 8102-8108,2010.

[11] W. Chen, S. Chen, H. Zhang and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, 2017, pp. 386-390.

[12] Wu H., Yang S., Huang Z., He J., Wang X.,"Type 2 diabetes mellitus prediction model based on data mining" in *Informatics in Medicine Unlocked*, Vol: 10,no.10 , pp. 100-107, 2018.

[13] 'Matlab for data science', [Online]. Available:https://analyticsindiamag.com/why-you-should-learn-matlab-for-data-science/

[14] H. A. Arnaldo and B. R. C. Bedregal, "A New Way to Obtain the Initial Centroid Clusters in Fuzzy C-Means Algorithm," *2013 2nd Workshop-School on Theoretical Computer Science*, Rio Grande, 2013, pp. 139-144.

[15] 'The logistic regression model', [Online]. Available: http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/logistic/LogisticRegression.pdf

[16] Ng, Andrew. "CS229 Lecture notes." CS229 Lecture notes 1.1 (2000): 1-3.