Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 2404- 2415
Research Article

Diagnosis of Diabetes Mellitus using Hybrid Techniques for Feature Selection and Classification

Ahmed Sami Jado'a [1], Prof. Dr. Ziyad Tariq Mustafa Al-Ta'i [2]

[1] Business Informatics College, University of Information Technology and Communications, Iraq
[2] Department of Computer science, Science College, Diyala University, Iraq
[1] ahmed.sami@uoitc.edu.iq, [2] ziyad1964tariq@uodiyala.edu.iq

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract
Diabetes is specified as the most chronic and deadliest disease that results in increasing blood sugar. The medical data mining approaches were utilized for detecting unobserved patterns in the medical fields of sets of data for medical diagnosis and treatment. Data classification for diabetes mellitus is quite significant. Two main measures are used in this analysis to differentiate between those with diabetes and others who are not. To find the most effective attributes for this disorder, the first move is to use Hybrid feature selection to find the most effective attributes. Classification is done in the second stage using "Logistic Regression and K-Nearest Neighbors algorithms". Where utilizing two types of data sets, the first is local, collected from consulting laboratories at Baquba General Hospital, and the second is global, which is the Pima India Diabetes Database (PIDD).The experiment on the Local dataset shows that LR without Hybrid feature gives an accuracy of 96%, while with Hybrid feature give accuracy of 98%, KNN without Hybrid feature give accuracy of 90%, while with Hybrid feature gives accuracy of 98%. The experiment on the Pima dataset shows that LR without Hybrid feature gives an accuracy of 76%, while with Hybrid feature gives accuracy of 90%, KNN without Hybrid feature gives accuracy of 81%, while with Hybrid feature gives accuracy of 85%.

Keywords: Diabetes Mellitus, Data mining, Diagnosis, Classification, Feature selection, Chi-square Test, Information gain, Hybrid feature, Logistic regression, K Nearest Neighbors.

1. Introduction

The method of removing latent information from large values of raw data is known as data mining. Data mining has been described as “the nontrivial extraction of previously unknown, implicit and potentially useful information from data". It is referred to as the science of extracting beneficial data from vast datasets. It is one of the activities involved in the database information discovery process[1].

Diabetes is expected to affect 9.3% of the global population (463 million people) in 2019, rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. Urban regions (10.8%) have a higher prevalence than rural areas (7.2%), and high-income countries (10.4%) have a higher prevalence than low-income countries (4.0 %). One-half of people with diabetes (50.1%) are unaware that they have the disease [2].
Due to the "International Diabetes Federation", diabetes affects 382 million individuals worldwide. By 2035, this figure would have risen to 592 million [3].
Data mining is the process of selecting, identifying, and showing vast volumes of data in order to reveal previously uncovered patterns or relationships that have a reliable and useful result for the data analyst. The KDD system is divided into several stages: Data mining includes measures such as data processing, data cleaning, data transfer, pattern matching (data mining), finding interpretation, finding definition, and finding evaluation [4].
Diabetes mellitus is a common disease where there is too much sugar (glucose) floating around in your blood. This occurs because either the pancreas can’t produce enough insulin or the cells in your body have become resistant to insulin. Diabetes affects the capability of the human body to utilize the energy present in food [5].

Diabetes is divided into three types: type 1, type 2, and gestational diabetes. If not well managed, diabetes may cause severe health problems. **Type1 Diabetes** Body does not able to produce insulin. Its affect children and young adults. Also it can affect at any age. Peoples affected by this type of diabetes to take insulin every day. **Type2 Diabetes** Body does not able to generate or utilize insulin. This type of diabetes mostly affected on middle aged and up in years. **Gestational Diabetes** Women's are mostly affected by this type of diabetes. During breastfeeding, this form of diabetes evolves. High blood sugar levels will impact your pregnancy and the health of your baby if you have gestational diabetes [6].

2. Related Work

Different related works were suggested in the field of diabetes, such as:
S.Selvakumar et.al. (2017)[7], Introduced a system for the prediction of diabetes mellitusbased on three classification methods, namely, "Logistic Regression, Multilayer Perceptron, and K-Nearest Neighbor". These classifiers give the accuracy of 69% for Logistic Regression, and 71% for Multilayer Perceptronand accuracy of80% for K-Nearest Neighbor.

DeeptiSisodia and Dilip Singh Sisodia (2018)[8], Developed a system that can perform prognosis of diabetes employing ranking algorithms. Which employing Naïve Bayes, SVM, and Decision Tree algorithms and compared to each other. The suggested approach for records' classification with the Naïve Bayes algorithm achieved 76.30 % accuracy, whereas the SVM algorithm achieved 65.10 % accuracy and with the Decision Tree achieved 73.82 % accuracy.

Tejas N. Joshi and Prof. Pramila M. Chawan (2018) [9], Proposed a system that can perform prognosis of diabetes employing Logistic Regression and SVM algorithms. These algorithms gave an accuracy of78% for Logistic Regression, and 79% for the SVM algorithm.

Seyed Ataaldin Mahmoudinejad Dezfuliet.al. (2019) [10], Developed an ensemble system utilizing data-mining approaches depended on four ranking approaches, namely, Simple tree classifier, Weighted KNN, Logistic regression, and Ensemble method algorithms to reveal human diabetes mellitus. These classifiers give an accuracy of 77 % for the Simple tree classifier, 77.3 % for Weighted KNN, an accuracy of 79.3 % for Logisticregression, andan accuracy of 80.60% for the Ensemble method.

Aswan Supriyadi Sunge et al. (2019)[11], Proposed a system use C4.5 algorithm to predication diabetes mellitus. The efficiency of the developed model evaluated by a database from the Pima India Diabetes dataset. These algorithms give the accuracy of 72.08%.

Nazim Razali et al. (2020)[12], proposed a system utilize various data mining mechanisms including "Naive Bayes, Sequential Minimal Optimization (SMO), RepTree and Simple Logistic Regression" for determining whether positive or negative consequence of diabetes diagnostic. The efficiency of the developed model evaluated by a database from the Pima India Diabetes dataset. These techniques give the accuracy of 73.60% for Naive Bayes, whereas give the accuracy of 75.70% for Simple Logistic Regression, give the accuracy of 75.10% for RepTree and give the accuracy of 74% for Sequential Minimal Optimization (SMO).

R. Kuppuchamy et.al. (2020) [13], Introduced a system to select significant attributes; it was done via the information gain method and C 4.5 classification algorithm. These algorithms gave an accuracy of 73.83 % withoutinformation gain and 74.87 % accuracy with information gain.

3. Standard Algorithms for Proposed System

This section contains the standard techniques that are used in hybrid techniques for selection and classification.

3.1. Feature Selection Algorithms

3.1.1. Chi-square Test

One of the function selection methodologies used in the filter method is the chi-squared test. The chi squared statistical test determines whether the two cases are independent. The mathematical freedom is denoted by the following equations (1) and (2) if X and Y are two cases[14].

$$P(XY) = P(X)\,P(Y) \quad (1)$$

Or

$$P(X/Y) = P(X) \text{ and } P(Y/X) = P(Y) \quad (2)$$

There is no connection between the cases, according to the null hypothesis. The designation is denoted by the events in the grouping. It's used in function discovery to see whether the appearance of a particular word and the appearance of a particular class are independent. As a result, the following quantities for each word are predicted, and they are ranked according to their score: (3) High scores on 2 mean that the null hypothesis of freedom (H0) can be refused, implying that the frequency of the word and class is contingent [14].

$$X_C^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

3.1.2. Information Gain

One of the function selection methodologies used in the filter method is Information Gain. Noise created by irrelevant functionality may be reduced using Information Gain. The functions with the most information dependent on a given class are detected by Information Gain. The best attributes are determined by first measuring the entropy value. Entropy is a chance event or attribute-based indicator of class uncertainty. The Entropy can be evaluated using equation (4) and then the Information Gain using equation (5) [15].

$$\text{Entropy (S)} = \sum_{i=1}^{c} (-p) log_2\ p_i \quad (4)$$

Here, c represents the total sum of meaning in the classification class, and Pi represents the percentage of S that corresponds to class i.

$$\text{Gain(S, A)} = \sum_{Values(A)} \frac{|S_v|}{S}\ Entropy\ (S) \quad (5)$$

Where A is an attribute, v is a potential value for attribute A, Values (A) is a set of potential values for A, $|S_v|$ is the number of samples for the value of v, | S | is the sum of all data samples and Entropy (S_v) is entropy for samples that have a value of v [15].

3.2. Classification Algorithms

3.2.1. Logistic Regression (LR)

The LR algorithm is a classification method. Provided a series of independent variables, it is used to predict a conditional result (1 / 0, Yes / No, True / False). Dummy variables are utilized to identify binary/categorical outcomes. Logistic regression is a type of linear regression in which the outcome variable is categorical and the dependent variable is the log of odds. In basic terms, it fits data to a logistic function to estimate the likelihood of an occurrence occurring. In equation (5) formula, $\mathbf{Pb_j}$ is the predicted probability by encoding it as 1, and **(1-** $\mathbf{Pb_j}$**)** is predicted probability by another decision and is encoded as 0 [16].

$$log\left(\frac{Pb_j}{1 - Pb_j}\right) = \alpha + \beta_1.X_{1j} + \beta_2.X_{2j} + \ldots + \beta_n.X_{nj} \quad (5)$$

Notation in Formula Logistics, wherein:

• α is the Intercept,

• $X_{1j} \ldots X_{nj}$ are independent attributes in the record –j,

• $\beta_1 \ldots \beta_n$ are slopes for independent attributes,

• n is the number of independent attributes,
• j is the number of records in the dataset.
To understanding the basics of the logistic regression classifier, let's start by reviewing the logistic function in Equation (6) [16].

$$P(t) = \frac{1}{1+e^{(-t)}} \quad (6)$$

3.2.2. K-Nearest Neighbors (KNN)

During the training process, KNN is a basic classification algorithm based on finding the nearest K neighbors. The distance between points is calculated using a similarity metric and the value of K, which shows the number of closest neighbors. KNN finds k neighbors for an undefined tuple by calculating the distance among data points in equation (7) formula using distance measurements such as Euclidean distance [17].

$$dist(X, Y) = \sqrt{\sum_{i=1}^{n}(xi, yi)^2} \quad (7)$$

4. Data Collection

4.1. Local Dataset

The information was gathered from Baqubah General Hospital's consulting labs in Iraq's Diyala Governorate. There are approximately 250 instances in this data collection. In the dataset, each person is identified by ten attributes. Gender, Age, HbA1C, Glucose, HDL Cholesterol, LDL Cholesterol, Total Cholesterol, Triglyceride, Creatinine, and Class are the attributes in the dataset. Both male and female genders were included in the data collection, which was at least 5 years old. The answer variable is conditional and takes the values 0 or 1, with 1 indicating a positive result for diabetes mellitus and 0 indicating a negative test. In class 1, there are 152 cases and in class 0, there are 98 cases.
Table (1) describes the collected local dataset.

Table (1): Description of Collected local Dataset

Att. No	Attribute	Description	Attribute Specification	Type
1	Gender	Gender for each person	1=Male 0=Female	Nominal
2	Age	Age for each person-years	5-74 years	Numeric
3	HbA1C	Hemoglobin A1C for each person 3-month plasma glucose concentration	4.80-5.90 %	Numeric
4	Glucose	Glucose levels in the blood for each person	4.11-6.05 mmol/L	Numeric
5	HDL Cholesterol	High-density lipoprotein cholesterol for each person	0.78-1.6 mmol/L	Numeric
6	LDL Cholesterol	Low density lipoprotein cholesterol for each person	0-2.6 mmol/L	Numeric
7	Total Cholesterol	Total amount of cholesterol in the blood	<5.2	Numeric
8	Triglyceride	Triglyceride for each person	0.86-1.9 mmol/L	Numeric

9	Creatinine	Creatinine for each person	62-106 µmol/L	Numeric
10	Class	Diagnosis of disease	1=True 0=False	Nominal

4.2 Global Data

The National Institute of Diabetes and Digestive and Kidney Diseases developed the Pima Indian Diabetic Database (PIDD). The information was gathered from Data World. Sets of data. According to open records, the whole patients are Pima Indian women who are at least 21 years old. There are a total of 768 cases, 268 of which are diabetic and 500 of which are not diabetic. The Pima class label categorized into (0 = False) indicates absence and (1 = True) presence of diabetes disease. Table (2) shows the Pima attributes description.

Table (2): Description of Pima Indian Diabetic Dataset

Att.No	Attribute	Description	Type
1	Pregnant	The number of years patients've been pregnant	Numeric
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
3	Blood pressure	The pressure of diastolic blood (mm Hg)	Numeric
4	Skin thickness	Triceps skin fold thickness (mm)	Numeric
5	Insulin	2-Hour serum insulin (mu U/ml)	Numeric
6	BMI	Body mass index (weight in kg/ (height in m) ^2)	Numeric
7	Diabetes pedigree function	The function of diabetes pedigree	Numeric
8	Age	Age (in years)	Numeric
9	Outcome	Class variable ((0=False) for tested negative for diabetes and ((1=True) for tested positive for diabetes)	Nominal

5. Proposed Model

The following ideas are presented in this paper: Preprocessing the data collection is the first approach. The second option is to reduce the amount of info. To increase the classification algorithm's accuracy, attribute subset selection techniques are used..

The suggested methodology in this paper is summarized in the diagram displayed in Fig. (1).

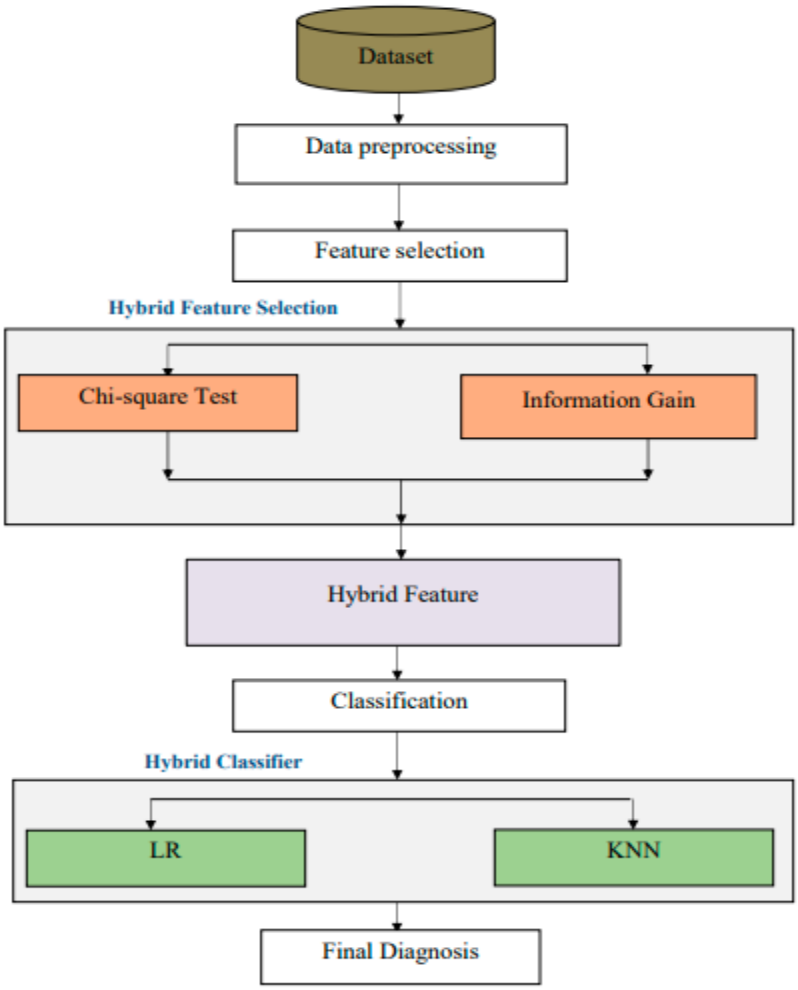

Figure (1): Block Diagram of the Proposed Model

5.1. Data Preprocessing

Preprocessing of data represents an essential initial step applied to raw data to prepare them for the analysis. Tools of analytic could give wrong results and be misled if impurities are included in the data like missing data. Hence, preprocess the data is essential before implementing the process of data analysis.

5.1.1. Data Cleaning

Data cleaning can be defined as the process that is utilized to ensure that data is clear and ensure that it is prepared for additional processing. Filling the missing data is a data cleaning process. There are many ways available to fill absent values such as removing records that have absent values or replacing them with casual values or substituting those absent values by the average value of the obtainable ones, which is the extreme advised approach to remove missing values from the used dataset.

5.1. Feature Normalization

This step is often adopted before the design of the classifier because it considers as a precaution when the feature values vary in different dynamic ranges. If normalization is not used, attributes with large values have a considerable impact on the design of the classifier. So the normalization role is to put all values within specific ranges features values are normalized by using the Min-Max method.

The initial data is transformed linearly by min-max normalization. Assume that $min_A$and max_A are the minimum and maximum values for attribute A, respectively. According to equation (8), min-max normalization maps a value v of A to v' in the range[new- min_A, new-max_A].

$$v_i^{`} = \frac{v_i - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new\_min_A \qquad (8)$$

Where $v_i^{`}$ represent features that normalized.

5.2. Selection of Feature

One of the most critical steps of preprocessing data mining is feature discovery, which involves selecting a subset of the initial feature spaces based on discrimination capabilities in order to increase data quality.

*** Hybrid Selection of Feature**

A Hybrid approach to decrease the number of features in the data set to a minimum to obtain the important and main features in the diagnosis by comparing the results of the two methods used in selecting the important features and then entering the results into the classification to obtain the best accuracy.

Algorithm (1): Hybrid selection using Chi-square test and Information gain
Input: Selected attributes using Chi-square test, Selected attributes using Information gain.
Output: Selected attributes
Begin **Step$_1$:** Make intersection between Selected attributes using Chi-square test and Selected attributes using Information gain. **Step$_2$:** Store as Selected attributes **End**

5.3. Classification

Since we have successfully identified the most appropriate features in the local dataset, global dataset (Pima), and handling all missing values based on pre-processing data and Hybrid feature selection process, the next step is to start the classification process using (LR and KNN) techniques as displayed in algorithm (2).

Algorithm (2): Classification Process using (LR and KNN) Techniques
Input: Local dataset, Global (PIMA) dataset
Output: Accuracy
Begin **Step$_1$:** Data Pre-processing. **Step$_2$:** Feature extraction by using the Hybrid Feature selection method. **Step$_3$:** Classify the dataset into training (80%) and testing (20%) dataset. **Step$_4$:** Apply the LR classifier, also the KNN classifier with **K** value is (11). **Step$_5$:** Diagnosis result for every classifier will be obtained. **Step$_6$:** The best diagnosis result will be used as a final diagnosis result. **End**

6. Evaluation Criteria

True positive (TP), true negative (TN), false positive (FP), and false-negative (FN) are the four relevant metrics used to determine the classification model throughout the uncertainty matrix (FN). Recall,

Precision, Accuracy, and F1-measure were used to deduce other performance metrics. These performance metrics are measured with the help of (TP, TN, FP, and FN). The following measures are used to test and compare classification models in this analysis [18].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$\text{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (12)$$

7. Results

Following the selection of the two datasets, they were subjected to two classification techniques (K-Nearest Neighbor and Logistic Regression) in order to determine which technique produces the best performance (accuracy) on the same dataset.

Evaluation results are given in table (3), table (4), and figure (2) on Local Dataset.

Table (3): Performance evaluation of LR and KNN without Hybrid feature

Model	Time (s)	Precision	Recall	F1-score	Accuracy	Confusion matrix	
LR	**0.04 sec**	**96%**	**96%**	**96%**	**96%**	20	1
						1	28
KNN	**0.03 sec**	**90%**	**90%**	**90%**	**90%**	19	2
						3	26

Table (4): Performance evaluation of LR and KNN with Hybrid feature

Model	Time (s)	Precision	Recall	F1- score	Accuracy	Confusion matrix	
LR	**0.02 sec**	**98%**	**97%**	**98%**	**98%**	19	1
						0	30
KNN	**0.02 sec**	**98%**	**98%**	**98%**	**98%**	25	0
						1	24

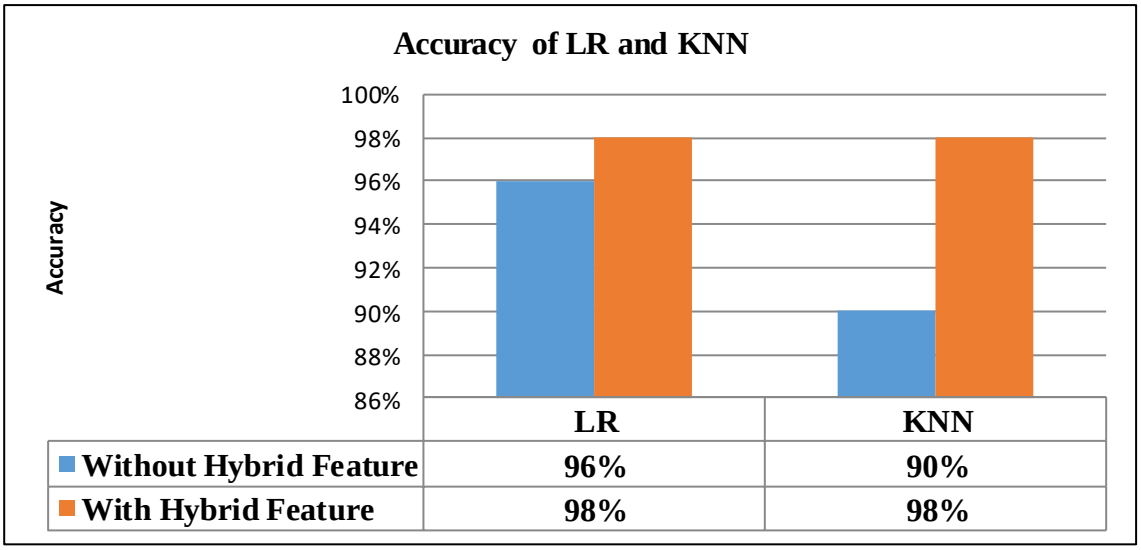

Figure (2): Accuracy of LR and KNN on Local dataset

Evaluation results are given in table (5), table (6), and figure (3) on Pima Dataset.

Table (5): Performance evaluation of LR and KNN without Hybrid feature

Model	Time (s)	Precision	Recall	F1- score	Accuracy	Confusion matrix	
LR	**0.05 sec**	**75%**	**71%**	**72%**	**76%**	40	5
						11	12
KNN	**0.03 sec**	**75%**	77%	76%	**81%**	43	8
						5	12

Table (6): Performance evaluation of LR and KNN with Hybrid feature

Model	Time (s)	Precision	Recall	F1- score	Accuracy	Confusion matrix	
LR	**0.03 sec**	**87%**	**85%**	**86%**	**90%**	48	3
						4	13
KNN	**0.02 sec**	**80%**	**80%**	**80%**	**85%**	46	5
						5	12

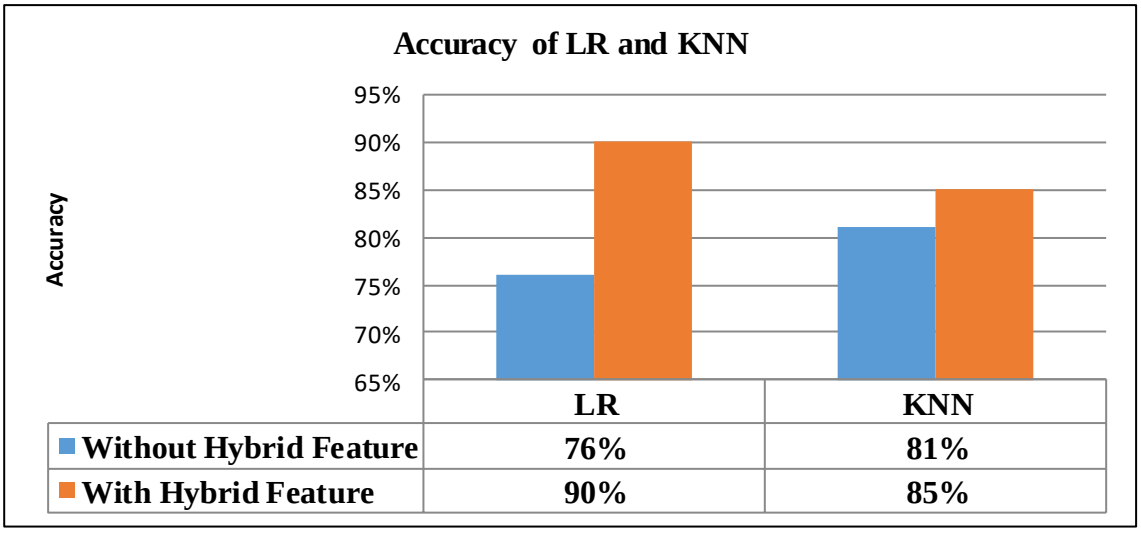

Figure (4): Accuracy of LR and KNN on Pima dataset

A comparison of the suggested approach with previous studies is displayed in table (7).

Table (7): Comparison of the Proposed Model Results with Other Studies

Author	Method	Classification accuracy
The proposed method	LR, KNN Without Hybrid feature on Local dataset	96%, 90%
	LR, KNN with Hybrid feature on Local dataset	98%, 98%
	LR, KNN Without Hybrid feature on Pima dataset	76%, 81%
	LR, KNN With Hybrid feature on Pima dataset	90%, 85
S.Selvakumar et.al. (2017) [7]	use of algorithms of classification Logistic Regression, Multilayer Perceptron, and K-Nearest Neighbor	69%,71%,80%
DeeptiSisodia and Dilip Singh Sisodia (2018) [8]	use Naïve Bayes , SVM, Decision Tree algorithms	76.30%, 65.10%,73.82%
Tejas N. Joshi and Prof. PramilaM. Chawan (2018) [9]	use of algorithms of classification Logistic Regression and SVM	78%, 79%
S. A. Mahmoudinejad Dezfuli et.al. (2019) [10]	use Simple tree classifier, Weighted KNN, Logistic regression and Ensemble method algorithms	77%, 77.3%, 79.3%, 80.60%
A. S. Sunge et.al. (2019)[11]	use C4.5 algorithm	72.08%
Nazim Razali et al. (2020) [12]	use Naive Bayes, Sequential Minimal Optimization (SMO), RepTree and Simple Logistic Regression	73.60%, 75.70% , 75.10% , 74%
R. Kuppuchamy et.al. (2020) [13]	use information gain feature	73.83, 74.87%

	selection method with C 4.5 classification algorithm	

8. Discussion and Conclusion

Diabetes mellitus is one of the world's widespread and complex diseases. By combining two datasets, the proposed model will diagnose diabetic and non-diabetic individuals. Result of classification on the Local dataset shows that LR without Hybrid feature gives an accuracy of 96% with time execution 0.04 sec, while with Hybrid feature give accuracy of 98% with time execution 0.02 sec, KNN without Hybrid feature give accuracy of 90% with time execution 0.03 sec, while with Hybrid feature gives accuracy of 98% with time execution 0.02 sec. On the Pima dataset shows that LR without Hybrid feature gives an accuracy of 76% with time execution 0.05 sec, while with Hybrid feature gives accuracy of 90% with time execution 0.03 sec, KNN without Hybrid feature gives accuracy of 81% with time execution 0.03 sec, while with Hybrid feature gives accuracy of 85% with time execution 0.02 sec. This proves that proposed hybrid techniques gives better performance than using the standard algorithms.

References

[1] P. Radha and B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," vol. 1, no. 6, pp. 334–339, 2014.

[2] P. Saeedi, I.Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107843, 2019, doi: 10.1016/j.diabres.2019.107843.

[3] M. S. F. Jebamalar and A. Anitha, "A Survey on Prediction of Dengue Fever Using Data Mining Techniques," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 2, no. 12, pp. 260–262, 2017.

[4] V. Mareeswari, R. Saranya, R. Mahalakshmi, and E. Preethi, "Prediction of diabetes using data mining techniques," *Res. J. Pharm. Technol.*, vol. 10, no. 4, pp. 1098–1104, 2017, doi: 10.5958/0974-360X.2017.00199.8.

[5] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

[6] M. Rajeswari and P. Prabhu, "A Review of Diabetic Prediction Using Machine Learning Techniques," *Int. J. Eng. Tech.*, vol. 5, no. 4, pp. 1–7, 2019.

[7] S. Selvakumar, K. S. Kannan, and S. Gothainachiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques," *Int. J. Stat. Syst.*, vol. 12, no. 2, pp. 183–188, 2017, [Online]. Available: http://www.ripublication.com.

[8] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[9] T. N. Joshi and P. M. Chawan, "Logistic Regression and SVM Based Diabetes Prediction System," *Int. J. Technol. Res. Eng.*, vol. 5, no. July, pp. 4347–4350, 2018, [Online]. Available: www.ijtre.com.

[10] S. A. Mahmoudinejad Dezfuli, S. R. Mahmoudinejad Dezfuli, S. V. Mahmoudinejad Dezfuli, and Y. Kiani, "Early Diagnosis of Diabetes Mellitus Using Data Mining and Classification Techniques," *Jundishapur J. Chronic Dis. Care*, vol. 8, no. 3, 2019, doi: 10.5812/jjcdc.94173.

[11] A. S. Sunge, H. L. H. S. Warnar, Y. Heryadi, E. Abdurachman, B. Soewito, and F. L. Gaol, "Prediction Diabetes Mellitus Using Decision Tree Models," *2019 Int. Congr. Appl. Inf. Technol. AIT 2019*, 2019, doi: 10.1109/AIT49014.2019.9144971.

[12] N. Razali, A. Mustapha, S. Z. S. Idrus, M. H. A. Wahab, and S. A. F. Madon, "Analyzing Diabetic Data using Classification," *J. Phys. Conf. Ser.*, vol. 1529, no. 2, 2020, doi: 10.1088/1742-6596/1529/2/022105.

[13] R. Kuppuchamy, T. Kamalavalli, S. Vinothini, N. Jayalakshmi, and N. Vallileka, "Correlation

based Ensemble Feature Selection Algorithm for Diagnosis of Diabetics," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 3S, pp. 373–373, 2020, doi: 10.35940/ijitee.c1080.0193s20.

[14] S. Vanaja and K. Ramesh Kumar, "Analysis of Feature Selection Algorithms on Classification: A Survey," *Int. J. Comput. Appl.*, vol. 96, no. 17, pp. 29–35, 2014, doi: 10.5120/16888-6910.

[15] A. S. Hassan, I. Malaserene, and A. A. Leema, "Diabetes Mellitus Prediction using Classification Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 2080–2084, 2020, doi: 10.35940/ijitee.e2692.039520.

[16] A. P. Wicaksono, T. Badriyah, and A. Basuki, "Comparison of The Data-Mining Methods in Predicting The Risk Level of Diabetes," *Emit. Int. J. Eng. Technol.*, vol. 4, no. 1, pp. 164–178, 2016, doi: 10.24003/emitter.v4i1.119.

[17] S. M. Gorade and P. A. Deo, "A Study Some Data Mining Classification Techniques," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 1, pp. 210–215, 2017, doi: 10.21884/ijmter.2017.4031.zt9tv.

[18] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," *Indian J. Sci. Technol.*, vol. 9, no. 43, 2016, doi: 10.17485/ijst/2016/v9i43/93874.