Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4815521, 9 pages
https://doi.org/10.1155/2022/4815521

Research Article

A Novel Architecture for Diabetes Patients' Prediction Using *K*-Means Clustering and SVM

Nitin Arora,[1] **Anupam Singh**,[2] **Mustafa Zuhaer Nayef Al-Dabagh**,[3] **and Sumit Kumar Maitra**[4]

[1]*Electronics & Computer Discipline, Indian Institute of Technology, Roorkee, India*
[2]*School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India*
[3]*Department of Computer Science, Knowledge University, Erbil 44001, Iraq*
[4]*Department of Electrical and Computer Science, Wachemo University, Hosaena, Ethiopia*

Correspondence should be addressed to Sumit Kumar Maitra; sumitmaitra@wcu.edu.et

Received 10 June 2022; Accepted 22 July 2022; Published 24 August 2022

Academic Editor: Punit Gupta

Diabetes is one of the alarming issues in today's era. It is a chronic disease that may cause many health-related problems. It is a group of syndrome that results in too much sugar in the blood. Diabetes's chronic hyperglycemia has been linked to long-term damage, organ breakdown, and organ failure, notably in the eyes, kidneys, nerves, heart, and veins. Machine learning has quickly advanced, and it is now used in many facets of medical health. The goal of this research is to create a model with the highest level of accuracy that can predict a patient's chance of developing diabetes. This paper proposes a novel architecture for predicting diabetes patients using the K-means clustering technique and support vector machine (SVM). The features extracted from K-means are then classified using an SVM classifier. A publicly available dataset, namely, the Pima Indians Diabetes Database, is tested using this approach. Accuracy of 98.7% is noted on the used dataset. On this dataset, the combined method performs better than the conventional SVM-based classification. This paper also compared the accuracy, precision, recall, and F1-score of the different machine learning techniques for classifying diabetes patients.

1. Introduction

Various forms of diabetes exist. In type 1, pancreatic insulin stops producing hormones. This hormone helps digest carbohydrates, fats, and proteins. In type 2 diabetes, cells associated with the digestive system cannot process insulin. Over time, the production of insulin in the body stops. Because of this, various internal organs start to become useless, which causes death. Type 3 diabetes is associated with pregnancy, in which a woman's blood sugar level increases [1–4]. A person cannot do anything special in advance for type 1 rescue. In type 2, nutritious food, regular exercise, and weight control are the best measures in prevention, and they can prevent 90% of diabetes [5, 6]. In 2014, in the clinic examination of 4.21 crore citizens, diabetes was found in 31 lakhs, i.e., about 7.75% of people. Awareness about diabetes is still low. 40% of affected citizens are unable to take proper care of themselves. They are not able to control their sugar level either. Half of them have never got an eye test done, even though they are at entire risk of retinopathy.

1.1. Motivation. According to the WHO, 7.30 million adults in India are in the grip of diabetes. 10.7 to 14.2% of the population is diabetic in cities, while 3 to 6.8% in villages. According to the central government, a quarter of India's 4.25 crore diabetes patients are in India. As per the National Family Health Survey-4, the number of diabetes patients in the country doubled in the year 2007 compared to 2007.3.5% of women and 3.5% of men aged 35 to 49 have diabetes. Diabetes is present in 9.7% of Indians over 80, while the same is present in 13.1% of people aged 60 to 69 and 13.2 in 70 to 79. After aging, the condition has reduced because the

patient is not survived. Among those who are left and are living for 60 years or even beyond, diabetes is about 36% or say one-third compared to people in the age group of 60 to 79. Heart disease, stroke, kidney disease, and blindness are also becoming a significant threat to health due to the increasing incidence of diabetes in the new generation [7–9].

1.2. Main Contributions. The contributions of this manuscript are as follows:

(i) This paper proposed a novel architecture to diagnose diabetes based on some parameters early.

(ii) This paper used *K*-means clustering combined with an SVM algorithm to classify the data.

(iii) For experiment purposes, the Pima Indians Diabetes Database is used. In the Pima Indians Diabetes Database, there are 668 female patients' data. 80% of these data are used to train machine and 20% to test the machine on the proposed architecture.

Sections break apart the remaining text of the paper: Section 2 describes the related work and compares the work done by all the researchers on the Pima Indians Diabetes Database using different techniques. Section 3 outlines the suggested approach. This section briefly describes our architecture and uses algorithms to improve the accuracy of predicting diabetes. Section 4 contains the experimental evaluation and results. This section briefly describes the used dataset with all the measured parameters. Section 5 of the paper puts it to a conclusion.

2. Related Work

Diabetes prediction using the Pima Indians Diabetes Database is a topic of interest among researchers during the last few decades. This section highlighted some of the methods used by the research to predict diabetes using the Pima Indians Diabetes Database and the accuracy achieved.

AlJarullah [10] has used the decision tree algorithm to predict type II diabetes. The data preprocessing part and the second diabetes prediction are completed in the first phase using the decision tree algorithm. The maximum accuracy achieved in this paper is 78.17%. Anand et al. [11] used higher-order neural network (HONN) combined with principal component analysis (PCA) to predict type II diabetes. In this paper, the authors used PCA to handle the missing data and also to scale the data in the same range of values. The maximum accuracy achieved in this paper is 89.47%. Banerjee et al. [12] used neural network, an evolutionary algorithm-based approach for predicting diabetes. This paper also compares the neural network model with other models. The maximum accuracy achieved in this paper is 93.5%. Barale and Shirke [13] used the K-means clustering algorithm combined with an artificial neural network (ANN) and K-means combined with logistic regression classifiers to predict diabetes. The maximum accuracy achieved in this paper is 98%. In order to uncover hidden patterns in the dataset, the K-means clustering technique is applied.

Chikh et al. [14] used the modified artificial immune recognition system (AIRS). In this, they used the fuzzy K-nearest neighbor algorithm to diagnose diabetes. The maximum accuracy achieved in this paper is 89.1%. Choubey and Paul[15] used the GA combined with multilayer perceptron neural network method for diagnosing diabetes. In the first phase, the genetic algorithm (GA) is used for feature selection, and in the second, diabetes classification is completed using multilayer perceptron neural network. The maximum accuracy achieved in this paper is 79.13%. Christobel and Sivaprakasam [16] used class-wise *K*-nearest neighbor (CkNN) to classify the diabetes dataset. In the first phase, data preprocessing is done, and the mean value is substituted in place of missing values. In the second phase of diabetes, classification is completed using modified KNN. The maximum accuracy achieved in this paper is 78.16%.

Deperlioğlu and Utku[17] used a multilayer feedforward NN structure trained by the Bayesian regularization algorithm and the mean square error function. The maximum accuracy achieved in this paper is 95.5%. In this study, the ANN was trained ten times. Gandhi and Prajapathi [18] used F-score, K-means clustering, Z-score normalization, and SVM. In the first phase, data preprocessing is done using F-score and K-means. In the second phase, diabetes classification is completed using SVM. The maximum accuracy achieved in this paper is 98%. Ganji and Abadeh [19] used ant colony optimization (ACO) to predict diabetes. The maximum accuracy achieved in this paper is 84.24%—using an ant colony-based classification method, a set of fuzzy rules for diabetic illness diagnosis may be extracted. Hayashi and Yukita [20] used Re-RX with J48 graft, combined with sampling selection techniques for predicting diabetes. As a "white-box" model, the recursive-rule extraction (Re-RX) method delivers extremely accurate categorization. The maximum accuracy achieved in this paper is 83.83%. Huang and Lu[21] used information gain (IG) along with DNN for the prediction of diabetes. The maximum accuracy achieved in this paper is 90.16%. Iyer et al. [22] used the J48 decision tree algorithm and naïve Bayes algorithm for the classification dataset and achieved an accuracy of 76.9% and 79.5%, respectively. Kahramanli and Allahverdhi[23] used an ANN and fuzzy NN to classify diabetes datasets. The maximum accuracy achieved in this paper is 86.8%.

Karatsiolis and Schizas[24] used a SVM with an RBF kernel and a SVM with a polynomial kernel to classify the diabetes dataset. First, the dataset was divided into two subsets. Then, one of the subsets, SVM with an RBF kernel, is applied, and on the other subset, SVM with a polynomial kernel is used. The maximum accuracy achieved in this paper is 82.2% and 81%, respectively. Karegowda et al. [25] used K-means clustering along with GA and CFS for the classification of the diabetes dataset. Classification accuracy of 96.68% is achieved in three phases. The K-means clustering algorithm is applied in the first phase to identify and eliminate incorrectly classified instances. In the second phase, a GA and correlation-based feature selection (CFS) are applied to extract relevant features. Finally, in the third phase, classification is done using K-nearest neighbor (KNN) algorithm. Karegowda et al. [26] used K-means

clustering combined with decision tree C4.5 to classify the diabetes dataset. In the first phase, K-means clustering is used to eliminate incorrect instances. In the second phase, the decision tree algorithm C4.5 is used to classify the data. The maximum accuracy achieved in this paper is 93.33%. Karegowda et al. [27] used a GA and back propagation network (BPN) to classify the data. The maximum accuracy achieved in this paper is 77.7%. Kayaer and Yildirim [28] used the GRNN to classify the data and achieved an accuracy of 80.21%.

Kumar Das et al. [29] used random forest and gradient boosting classifiers to classify diabetes datasets and achieved an accuracy of 90%. Initially, data preprocessing is done, and then the classifier is applied to classify the data. Senthil Kumar et al. [30] used covering-based rough set classification for the dataset classification. This is a pattern-based approach. Maximum accuracy of 79.34% is achieved using this procedure. Kumari and Chitra[31] used SVM with RBF kernel to classify the data and achieved an accuracy of 75.5%. Nirmala Devi et al. [32] used amalgam KNN to classify the data and achieved an accuracy of 97.4%. This amalgam of KNN consists of K-means with KNN. K-means algorithm is used to identify missing values. Missing values are replaced by the mean and median in this algorithm. Patil et al. [33] used K-means clustering combined with decision tree C4.5 and achieved an accuracy of 92.38% to classify the dataset. Polat [34] used fuzzy C-means combined with SVM and KNN and weighting methods (FCMAW) and achieved an accuracy of 91.41 and 84.38, respectively. Polat et al. [35] used GDA and least square support vector and achieved an accuracy of 82.05% to classify the data. Rado et al. [36] used random forest combined with recursive feature elimination, and the accuracy achieved was 73%. Raghavendraet al. [37] used a neural network model with a backward elimination feature selection method and made the accuracy of 84.52% to classify the dataset. Rajni and Amandeep [38] achieved a classification accuracy of 72.9% by using the RB-Bayes algorithm. In this, the mean is used to handle the missing values. Ramana and Boddu [39] used the naïve Bayes classification algorithm, and the accuracy achieved was 76.34%. Balajiet al. [40] used a deep NN restricted Boltzmann machine, and 80.9% accuracy was achieved.

Vaishali et al. [41] used Goldberg's GA combined with a multi-objective evolutionary fuzzy classifier to classify the type 2 diabetes dataset. In the first stage, essential features are extracted using Goldberg's GA. In the second stage, the multi-objective evolutionary fuzzy classifier is applied, and an accuracy of 83.04% is achieved using this method. Vosoulipour et al. [42] used NN and ANFIS structures and achieved an accuracy of 81.3%. Wong and Lease [43] used Cartesian genetic programming and achieved an accuracy of 80.5%. Wu et al. [44] used an improved K-means algorithm and the logistic regression algorithm for the dataset classification and achieved an accuracy of 95.42%. Zolfaghari [45] used SVM combined with NN and achieved an accuracy of 88.04%, and Bano and Khan[46] used K-NN and achieved an accuracy of 82.29% to classify the dataset.

An automated model for diagnosing diabetes was reported by Lakhvaniet al. [47] utilizing a three-layered artificial neural network (ANN). For neuron activation, the authors employed a logistic activation function, and they trained the model using the quasi-Newton approach. Through the use of the Pima Indian Diabetes Dataset, Patil and Ingle[48] offered a comparative analysis of different ML classification algorithms with diabetes prediction. For statistical modeling and accuracy verification, authors employed KNN, LR, which is based on the regression problem, naive Bayes probabilistic classifier, SVM with both linear and nonlinear kernel, and decision tree with RF classifier. 80.20 percent accuracy is the highest possible. LDA and GA were employed for feature selection by Alharan et al. [49] to increase the classification accuracy for diabetes. The approach has a maximum accuracy of 90.89 percent. Sivaranjani et al. [50] presented a model for diabetes categorization using SVM and RF techniques. PCA is also used to reduce the number of dimensions, with maximum accuracy rates of 83 and 81.4 percent, respectively.

In all the techniques used by the researchers, the main challenge is to improve the accuracy of the system for early diagnosis of diabetes. To overcome this problem, this paper suggested a fusion technique in two phases. In the first phase, data preprocessing is done using K-means, and in the second phase, diabetes classification is completed using SVM to achieve the maximum accuracy. Techniques used by different researchers and achieved accuracy are summarized in Table 1.

3. Proposed Methodology

This section describes the proposed Pima diabetes patient classification model using K-means clustering and SVM. Figure 1 presents an overview of the suggested model. The proposed model first created the clusters using the K-means clustering and then used the SVM for the classification.

3.1. K-Means Clustering Algorithm. K-Means algorithm is used to cluster the dataset into different classes. K-Means works for multi-dimensional data. For two-dimensional data, the example is shown in Figure 2.

The following steps are used in the K-means clustering algorithm [51]:

(1) Choose the K number of clusters.

(2) Choose at random k points. These k points will be the centroids of the k clusters. It is not necessarily that these k points are from dataset. Any k points can be selected.

(3) Assign each data point to the nearest centroid, and the resulting k cluster will be formed. The Euclidian distance is used to calculate distance.

(4) Determine and set each cluster's new centroid.

(5) Change the centroid that corresponds to each data point. If there was a reassignment, proceed to step 4, otherwise, end.

Table 1: Comparative study of existing approaches used by the researchers and accuracy achieved.

Sr. no.	Method used	Accuracy (%)	Reference
1	Decision tree	78.17	[10]
2	Higher-order NN with PCS	89.47	[11]
3	NN	93.5	[12]
4	Classifier using the K-means algorithm with logistic regression	98	[13]
5	Fuzzy K-nearest neighbors	89.1	[14]
6	GA combined with multilayer perceptron neural network	79.13	[15]
7	Class-wise *K*-nearest neighbor (CkNN)	78.16	[16]
8	Multilayer feedforward neural network	95.5	[17]
9	F-score, K-means clustering along with Z-score normalization and SVM	98	[18]
10	Ant colony optimization (ACO)	84.24	[19]
11	Re-RX with J48 graft, combined with sampling selection techniques	83.83	[20]
12	Information gain (IG) along with deep NN	90.26	[21]
13	Decision tree and naïve Bayes	76.9 and 79.5 respectively	[22]
14	ANN and FNN	86.8	[23]
15	SVM with an RBF kernel and with a polynomial kernel	82.2	[24]
16	K-means clustering along with GA and CFS	96.68	[25]
17	K-means clustering combined with decision tree C4.5	93.33	[26]
18	GA and back propagation network (BPN)	77.7	[27]
19	General regression neural network (GRNN)	80.21	[28]
20	Random forest and gradient boosting classifiers	90	[29]
21	Covering-based rough set	79.34	[30]
22	SVM (with RBF kernel)	75.5	[31]
23	Amalgam KNN	97.4	[32]
24	K-means clustering combined with decision tree C4.5	92.38	[33]
25	Fuzzy C-means combined with SVM and KNN and weighting methods (FCMAW)	91.41 and 84.38, respectively	[34]
26	GDA and least square support vector	82.05	[35]
27	Random forest combined with recursive feature elimination	73	[36]
28	Neural network model with backward elimination feature selection method	84.52	[37]
29	RB-Bayes	72.9	[38]
30	Naïve Bayes	76.3	[39]
31	Deep neural network restricted Boltzmann machine	80.9	[40]
32	Goldberg's GA combined with multi-objective evolutionary fuzzy classifier	83.04	[41]
33	Neural network and ANFIS structures	81.3	[42]
34	Cartesian genetic programming	80.5	[43]
35	Improved the K-means and the logistic regression	95.42	[44]
36	SVM combined with neural network	88.04	[45]
37	KNN	82.29	[46]

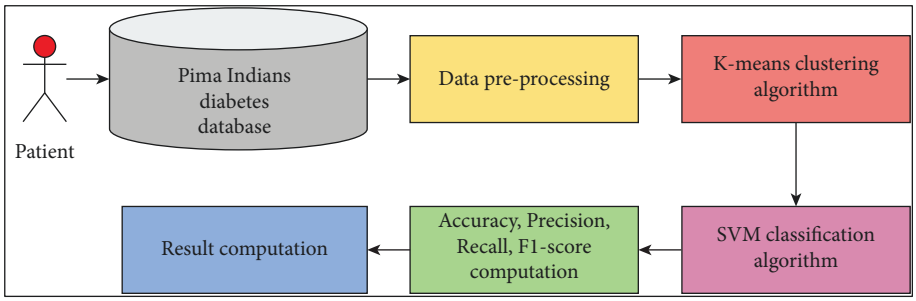

Figure 1: The suggested Pima diabetic patient classification model utilizing K-means clustering and SVM.

The number of clusters (in step 1) is computed using the elbow method. For the used dataset, the number of clusters is 5.

3.2. SVM Classification Algorithm. SVM was developed nationally in the 1960s and later found in the 1990s. SVM is very popular in machine learning because SVM is a robust algorithm. SVM is very different from other machine learning algorithms. SVM is about finding the best decision boundary that helps to separate the dataset into different classes. SVM separates the types through the maximum margin boundary between support vectors. For the best boundary, the sum of the distances of the boundary line from support vectors should be maximum. This boundary line is known as the maximum

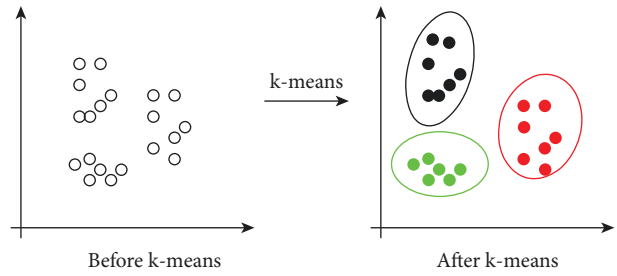

Figure 2: Data space before and after applying K-means clustering algorithm.

margin hyperplane or maximum margin classifier. The support vector classification (SVC) class is used to implement SVM. In SVC, there are many parameters. In this model, we used the linear kernel for linear classification [52].

4. Experimental Evaluation and Results

The specifics of the dataset utilized in this investigation are presented in this section. Results are calculated using various categorization algorithms and suggested methods. The details are as follows.

4.1. Dataset Description. This paper used a publicly available dataset, namely, Pima Indians Diabetes Dataset [53]. This dataset contains the data of a total of 668 female patients with eight independent parameters, namely, pregnancies, glucose, blood pressure (BP), skin thickness (ST), insulin, BMI, diabetes pedigree function and age, and one dependent parameter, outcome [53]. The first five records of the dataset are presented in Table 2.

All the parameters of the used dataset are as follows:

(i) Pregnancies: This parameter represents the number of times pregnant.

(ii) Glucose: During an oral glucose tolerance test, plasma glucose concentration exceeded 2 hours.

(iii) Blood Pressure: Diastolic heart rate (mm Hg).

(iv) Skin Thickness: This parameter represents the triceps skinfold thickness (mm).

(v) Insulin: This parameter expresses the 2-hour serum insulin (mu U/ml).

(vi) BMI: This parameter describes the body mass index (weight in kg/height in $m)^2$.

(vii) Diabetes Pedigree Function (DPF): DPF is a function that scores the likelihood of diabetes based on family history.

(viii) Age: This parameter represents the age of the person (in years).

(ix) Outcome: This parameter represents the class variable. 0 means nondiabetic, and 1 means diabetic.

4.2. Performance Measure. All the approaches are compared using accuracy, precision, recall, and F1-score for performance measures. Accuracy, precision, recall, and F1-score are computed using (1)–(4) [54]. The confusion matrix is plotted for calculating all the performance measure parameters. The generated confusion matrix of the proposed method is shown in Figure 3.

The used performance measurement parameters are for confusion matrix (Table 3) with two classes (binary classification).

(i) Accuracy: The proportion of correct classification (true positive and true negative) from the overall number of cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

(ii) Precision: The percentage of instances that were correctly classified as positive (true positive) when they were expected to be positive.

$$\text{Precision}\,(P) = \frac{TP}{TP + FP}. \tag{2}$$

(iii) Recall: The percentage of correctly classifying positive instances as positive (true positive).

$$\text{Recall}\,(R) = \frac{TP}{TP + FN}. \tag{3}$$

(iv) F1-score: The balance between precision and recall is shown by the F1-score.

$$F1 - \text{score} = \frac{2 \times P \times R}{P + R}. \tag{4}$$

4.3. Experimental Results. The discussion presented in Table 4 shows the result of the proposed approach to the Pima Indians Diabetes Dataset. The accuracy of 98.7% is recorded using the proposed method, whereas the accuracy of 82.46% is recorded using only the SVM classification algorithm. An improvement of 19.69% is achieved on the Pima Indians Diabetes Dataset.

The comparison of the various classification methods, namely, decision tree, random forest, kernel SVM, naive Bayes, KNN, logistic regression, SVM, and the proposed approach based with respect to accuracy, precision, recall, and F1-score, is shown in Figures 4–7.

Table 2: First five records in the Pima Indians Diabetes Dataset.

Sr. no.	Pregnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

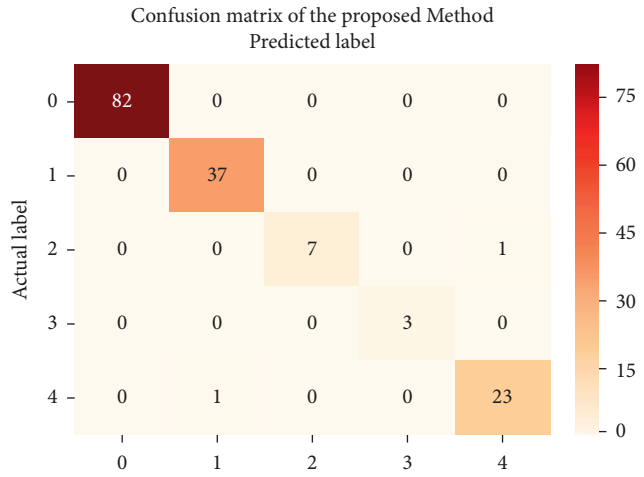

Figure 3: Confusion matrix of the proposed method.

Table 3: Confusion matrix for two classes.

Actual class	Predicted class	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

TN: true negative, TP: true positive, FP: false positive, FN: false negative.

Table 4: Performance evaluation of several classification methods.

Method used	Accuracy	Precision	Recall	F1-score
Decision tree	70.77	67	69	67.5
Random forest	78.57	75	72.5	73.5
Kernel SVM	79.22	76	72.5	74
Naive Bayes	79.22	75.5	74.5	74.5
KNN	79.87	76.5	75.5	75.5
Logistic regression	82.46	80	77	78
SVM	82.46	80	77	78
Proposed approach	**98.7**	**98.6**	**96.8**	**97.5**

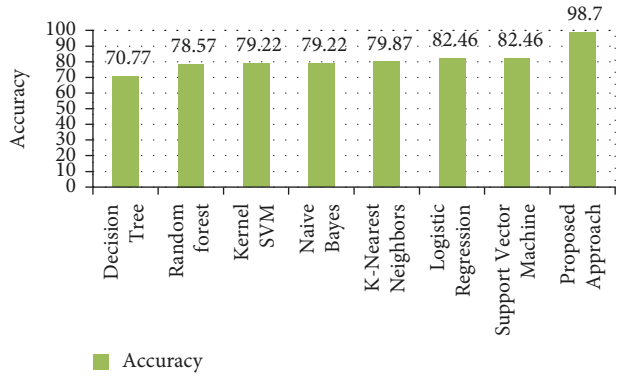

Figure 4: Accuracy using different classification approaches.

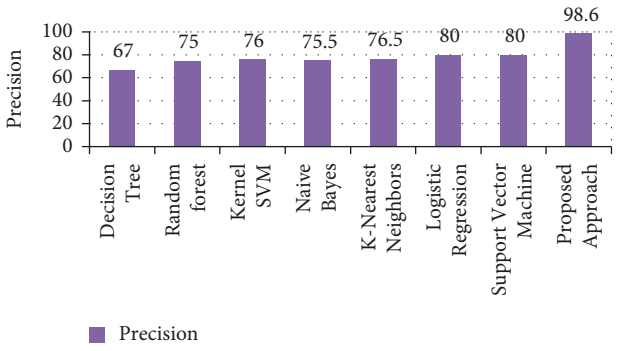

Figure 5: Precision using different classification approaches.

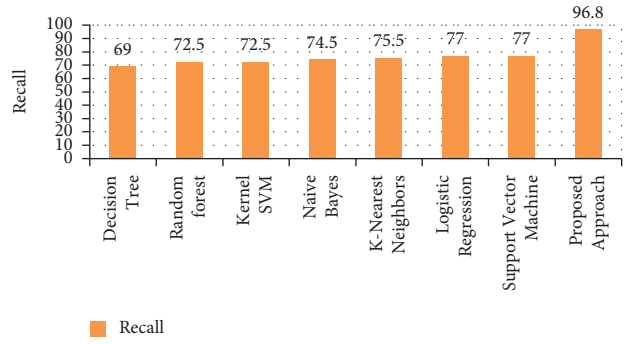

Figure 6: Recall using different classification approaches.

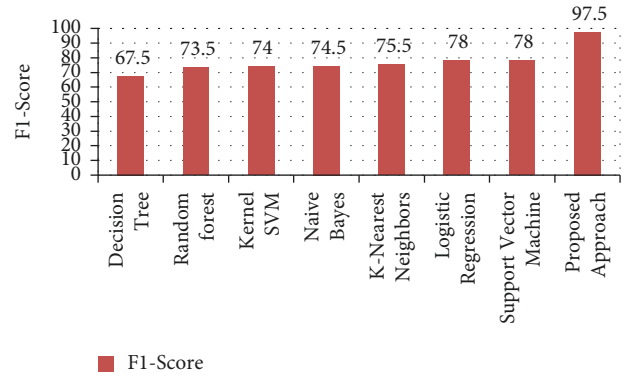

Figure 7: F1-score using different classification approaches.

The accuracy, precision, recall, and F1-score of the proposed method are 98.7%, 98.6%, 96.8%, and 97.5%, respectively.

5. Conclusion and Future Scope

This study suggested a brand-new architecture for diabetes patient categorization using K-means clustering and SVM. The clusters of the database are designed using a K-means clustering method. The predictions are then computed based on the created clusters considered as features for categorization using SVM. The Pima Indians Diabetes Database is used to verify the approach's resilience against a publicly accessible dataset. The Pima Indians Diabetes Database has 668 female patients' data. 80% of these data are used to train machine and 20% to test the machine on the proposed architecture, with a maximum accuracy of 98.7%. By obtaining more reliable characteristics from the database, the classification rates may rise in the future. Additionally, combining techniques like decision fusion of several classifiers might help the classification process.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Spranger, A. Kroke, M. Möhlig et al., "Adiponectin and protection against type 2 diabetes mellitus," *The Lancet*, vol. 361, no. 9353, pp. 226–228, 2003.

[2] B. B. Lowell and G. I. Shulman, "Mitochondrial dysfunction and type 2 diabetes," *Science*, vol. 307, no. 5708, pp. 384–387, 2005.

[3] R. S. Lindsay, T. Funahashi, R. L. Hanson et al., "Adiponectin and development of type 2 diabetes in the Pima Indian population," *The Lancet*, vol. 360, no. 9326, pp. 57-58, 2002.

[4] M. E. Miller, R. P. Byington, D. C. Goff Jr et al., "Effects of intensive glucose lowering in type 2 diabetes," *New England Journal of Medicine*, vol. 358, no. 24, pp. 2545–2559, 2008.

[5] G. Klöppel, M. Löhr, K. Habich, M. Oberholzer, and P. U. Heitz, "Islet pathology and the pathogenesis of type 1 and type 2 diabetes mellitus revisited," *Pathology & Immunopathology Research*, vol. 4, no. 2, pp. 110–125, 1985.

[6] B. Zinman, C. Wanner, J. M. Lachin et al., "Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes," *New England Journal of Medicine*, vol. 373, no. 22, pp. 2117–2128, 2015.

[7] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 32, no. Suppl 1, pp. S62–S67, 2009, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613584/.

[8] Health Information, "Diabetes in older people," *Health Information*, https://www.nia.nih.gov/health/diabetes-older-people, 2020.

[9] An overview of diabetes types and treatments, "An overview of diabetes types and treatments," *RECALL OF METFORMIN EXTENDED RELEASE*, https://www.medicalnewstoday.com/articles/323627.php, 2020.

[10] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Proceedings of the 2011 International conference on innovations in information technology*, pp. 303–307, IEEE, 2011, April.

[11] R. Anand, V. P. S. Kirar, and K. Burse, "K-fold cross-validation and classification accuracy of PIMA Indian diabetes data set using higher-order neural networks and PCA," *International Journal of Soft Computing and Engineering*, vol. 2, no. 6, pp. 2231–2307, 2013.

[12] C. Banerjee, S. Paul, and M. Ghoshal, "An evolutionary algorithm based parameter estimation using pima Indians diabetes dataset," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 6, pp. 374–377, 2017.

[13] M. S. Barale and D. T. Shirke, "Cascaded modeling for PIMA Indian diabetes data," *International Journal of Computer Application*, vol. 139, no. 11, pp. 1–4, 2016.

[14] M. A. Chikh, M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an artificial immune recognition system2 (AIRS2) with fuzzy k-nearest neighbor," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2721–2729, 2012.

[15] D. K. Choubey and S. Paul, "GA_MLP NN: a hybrid intelligent system for diabetes disease diagnosis," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 1, pp. 49–59, 2016.

[16] Y. A. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset," *International Journal of Engineering and Advanced Technology*, vol. 2, no. 3, pp. 396–200, 2013.

[17] O. Deperlioglu and K. O. S. E. Utku, "Diabetes determination using retraining neural network," in *Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp. 1–5, IEEE, 2018, September.

[18] K. K. Gandhi and N. B. Prajapati, "Diabetes prediction using feature selection and classification," *International journal of advance Engineering and Research Development*, vol. 1, no. 05, 2014.

[19] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14650–14659, 2011.

[20] Y. Hayashi and S. Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset," *Informatics in Medicine Unlocked*, vol. 2, pp. 92–104, 2016.

[21] L. Huang and C. Lu, "Intelligent diagnosis of diabetes based on information Gain and deep neural network," in *Proceedings of the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 493–496, IEEE, 2018, November.

[22] A. Iyer, S. Jeyalatha, and R. Sumbaly, *Diagnosis of Diabetes Using Classification Mining Techniques*, http://arXiv.org/abs/1502.03774, 2015.

[23] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 82–89, 2008.

[24] S. Karatsiolis and C. N. Schizas, "Region-based support vector machine algorithm for medical diagnosis on the pima Indian diabetes dataset," in *Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, pp. 139–144, IEEE, 2012, November.

[25] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology*, vol. 1, no. 3, pp. 147–151, 2012.

[26] A. G. Karegowda, V. Punya, M. A. Jayaram, and A. S. Manjunath, "Rule-based classification for diabetic patients using cascaded k-means and decision tree C4. 5," *International Journal of Computer Application*, vol. 45, no. 12, pp. 45–50, 2012.

[27] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Application of genetic Algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes," *International Journal of Soft Computing*, vol. 2, no. 2, pp. 15–23, 2011.

[28] K. Kayaer and T. Yıldırım, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," in *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, vol. 181, p. 184p. 184, 2003, June.

[29] S. Kumar Das, A. Kumar Mishra, and P. Roy, "Automatic diabetes prediction using tree-based ensemble learners," *International Journal of Computational Intelligence & IoT*, vol. 2, no. 2, 2019.

[30] S. Senthil Kumar, H. Hannah Inbarani, A. T. Azar, and K. Polat, "Covering-based rough set classification system," *Neural Computing & Applications*, vol. 28, no. 10, pp. 2879–2888, 2017.

[31] V. A. Kumari and R. Chitra, "Classification of diabetes disease using a support vector machine," *International Journal of Engineering Research in Africa*, vol. 3, no. 2, pp. 1797–1801, 2013.

[32] M. NirmalaDevi, S. A. alias Balamurugan, and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," in *Proceedings of the 2013 IEEE International Conference ON Emerging Trends in Computing, Communication, and Nanotechnology (ICECCN)*, pp. 691–695, IEEE, 2013, March.

[33] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for type-2 diabetic patients," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8102–8108, 2010.

[34] K. Polat, "Intelligent recognition of diabetes disease via FCM based attribute weighting," *International Journal of Computer and Information Engineering*, vol. 10, no. 4, pp. 783–787, 2016.

[35] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine," *Expert Systems with Applications*, vol. 34, no. 1, pp. 482–487, 2008.

[36] O. Rado, N. Ali, H. M. Sani, A. Idris, and D. Neagu, "Performance analysis of feature selection methods for classification of healthcare datasets," in *Proceedings of the Intelligent Computing-Proceedings of the Computing Conference*, pp. 929–938, Springer, Cham, 2019, July.

[37] S. Raghavendra, S. Kumar, and B. K. Raghavendra, "Evaluating the performance of neural network using feature selection methods on pima INDIAN diabetes dataset," *Journal of emerging technologies and innovative research*, 2018.

[38] R. Rajni and A. Amandeep, "RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, p. 4866, 2019.

[39] B. V. Ramana and R. S. K. Boddu, "Performance comparison of classification algorithms on medical datasets," in *Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0140–0145, IEEE, 2019, January.

[40] H. Balaji, N. Iyengar, R. D. Caytiles, and R. D. Caytiles, "Optimal predictive analytics of Pima diabetics using deep learning," *International Journal of Database Theory and Application*, vol. 10, no. 9, pp. 47–62, 2017.

[41] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in *Proceedings of the 2017 International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1–5, IEEE, 2017, October.

[42] A. Vosoulipour, M. Teshnehlab, and H. A. Moghadam, "Classification of diabetes mellitus dataset based-on artificial neural networks and ANFIS," in *Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering 2008 Berlin, Heidelberg*, pp. 27–30, Springer, 2008.

[43] W. K. Wong and B. A. Lease, "Spherical bounding classifier using CGP generated transforms," in *Proceedings of the IOP Conference Series: Materials Science and Engineering*, vol. 495, no. No. 1, p. 012016p. 012016, 2019, April.

[44] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.

[45] R. Zolfaghari, "Diagnosis of diabetes in female population of Pima Indian heritage with an ensemble of bp neural network

and SVM," *Int. J. Comput. Eng. Manag*, vol. 15, pp. 2230–7893, 2012.

[46] S. Bano and M. N. A. Khan, "A framework to improve diabetes prediction using k-NN and SVM," *International Journal of Computer Science and Information Security*, vol. 14, no. 11, p. 450, 2016.

[47] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele, and D. Somwanshi, "Prediction of the onset of diabetes using artificial neural network and pima indians diabetes dataset," in *Proceedings of the 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–6, IEEE, 2020, December.

[48] V. Patil and D. R. Ingle, "Comparative analysis of different ML classification algorithms with diabetes prediction through Pima Indian diabetics dataset," in *Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT)*, pp. 1–9, IEEE, 2021, June.

[49] A. F. Alharan, Z. M. Algelal, N. S. Ali, and N. Al-Garaawi, "Improving classification performance for diabetes with linear discriminant analysis and genetic algorithm," in *Proceedings of the 2021 Palestinian International Conference on Information and Communication Technology (PICICT)*, pp. 38–44, IEEE, 2021, September.

[50] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction," in *Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 141–146, IEEE, 2021, March.

[51] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: a k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, p. 100, 1979.

[52] R. Ranjan, A. Singh, A. Rizvi, and T. Srivastava, "Classification of chest diseases using convolutional neural network," in *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019). Lecture Notes in Networks and Systems*, P. Singh, W. Pawłowski, S. Tanwar, N. Kumar, J. Rodrigues, and M. Obaidat, Eds., vol. 121, Singapore, Springer, 2020.

[53] J. W. Smith and J. E. Everhart, "Predict the onset of diabetes based on diagnostic measures," *Pima Indians Diabetes Database*, https://www.kaggle.com/uciml/pima-indians-diabetes-database/download, 1988.

[54] T. N. Do, P. Lenca, S. Lallich, and N. K. Pham, "Classifying very-high-dimensional data with random forests of oblique decision trees," in *Advances in Knowledge Discovery and Management*, pp. 39–55, Springer, Berlin, Heidelberg, 2010.