VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2023, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 11, Number 4, October-December 2023 pp:32-44

Enhanced Diabetic Prediction Using Fuzzy C-Means Preprocessing and Random Forest Ensemble Learning

Priha Bhatti [1,2*], **Khalid Mahboob** [2], **S. Saad Naeem** [3], **Iqra Heer Bhatti** [4], **Noorulain Kamran** [5]

[1]PhD Scholar, Department of Computer Science, Muhammad Ali Jinnah University, Karachi, Pakistan; [2]Department of Software Engineering, Sir Syed University of Engineering Technology, Karachi, Pakistan; [3] Chief Executive Officer ,Xicom Solution Pvd.Ltd, Karachi, Pakistan; [4]Department of General Surgery, Ziauddin University Hospital; [5]Department of Pharmacology, Dow University of Health Sciences

Keywords: *Data Mining, PCA, Fuzzy C-Means, and Random Forest Algorithm*

Journal Info:
Submitted:
October 15, 2023
Accepted:
November 25, 2023
Published:
December 02, 2023

Abstract

Diabetes claims the lives of thousands each year, and many individuals remain oblivious to their condition until it reaches a critical stage. This study presents a data mining-based approach aimed at enhancing the early detection and prediction of diabetes, utilizing data from the Pima Indian Diabetes dataset. Despite the adaptability of fuzzy C-Means for various data types, the ultimate outcome of the clustering process hinges on the initial placement of cluster centers. Additionally, precision in data clustering is crucial; it can furnish either extensive, well-grouped data for the random forest or limited data, constraining its efficacy. Our principal objective was to enhance the accuracy of fuzzy C-means clustering and the random forest. To boost the model's performance, we incorporated PCA, fuzzy c-means, and the Random Forest approach. Various algorithmic combinations were employed, and the results unequivocally demonstrate that our model surpasses the original outcomes of the Pima Indian Diabetes Dataset in terms of accuracy. The diabetic prediction model achieved a remarkable accuracy of 97.40% through the utilization of PCA, logistic regression, and K-Means. However, when employing PCA in conjunction with fuzzy C-means and random forests, an even higher accuracy of 98.96% was attained. Empirical evidence confirms that the implementation of PCA significantly enhanced the accuracy of both the fuzzy C-means clustering approach and the random forest classifier, deviating from previous findings.

***Correspondence author email address:** FA23PHCS0001@maju.edu.pk
DOI: 10.21015/vtse.v11i4.1657

1 Introduction

In the landscape of global health challenges, diabetes emerges as a formidable and widespread non-communicable disease, poised to exact a staggering toll on human lives. Projections indicate that by 2040, diabetes is expected to become the leading cause of death, affecting an estimated 642 million individuals worldwide. Characterized by elevated blood glucose levels, this complex condition manifests when the body's ability to produce insulin or use it effectively falters.

Diabetes, a globally recognized non-communicable disease, is projected to be the leading cause of death by 2040 [8, 13]. Machine Learning and artificial intelligence, sub-disciplines of computer science [20, 21], focus on creating intelligent machines capable of problem-solving, learning, thinking, and performing human-like tasks [22]. AI finds applications in various industries, including manufacturing, weather forecasting, medical care, account management, website development, and climate change mitigation [23, 24]. Machine learning, already widely used, includes applications like face and voice recognition systems, biometrics, robotic surgery, and remote sensing. In the healthcare industry, AI and machine learning play crucial roles [25, 26], with ML techniques frequently applied in diabetes prediction, yielding improved results [27].

Diabetes occurs when blood glucose levels are abnormally high, also known as blood sugar. Blood glucose, derived from the diet, serves as the primary source of sugar. The pancreas produces insulin, a hormone facilitating the absorption of glucose by cells for use as fuel. Insufficient insulin production or ineffective utilization can lead to diabetes.

Machine learning algorithms find application in various medical fields, as evidenced by studies utilizing techniques, algorithms, and data mining to develop diabetic prediction models. The healthcare sector, with physicians and scientists dedicated to improving the system [28, 29], continues ongoing research in diagnosing and predicting diabetes, emphasizing the need for enhanced prediction models [30].

Diabetes is the most prevalent disease affecting humans, resulting from inadequate insulin synthesis and elevated blood sugar levels. Detecting signs and symptoms is crucial before clinical examination, but despite easy access to these symptoms, accurate diabetes prediction remains a significant challenge. Researchers focus on accurately diagnosing diabetic conditions by collecting substantial data. Current standard stages for recognizing diabetes utilize minimal processes, yet fall short of achieving the highest possible detection accuracy [31].

Early detection and symptomatic treatment are crucial for the well-being of pre-diabetic patients. An intelligent medical diagnosis system based on symptoms, signs, laboratory tests, and observations can aid in disease detection and prevention. Artificial Intelligence (AI) has been applied to medical diagnosis systems in various ways for disease detection [32].

This exploration employs machine learning algorithms to enhance the predictive accuracy of diabetes diagnosis. The process begins with data accession from a designated database, followed by thorough data medication to prepare the dataset for analysis. Data preprocessing and standardization then upgrade the dataset for subsequent logical procedures. Principal Component Analysis (PCA) is utilized to homogenize the data, reducing implicit biases and ensuring uniformity in subsequent analyses. Fuzzy C-Means Clustering, integrated into the exploration, employs a supervised bracket approach based on insights from PCA. This binary-rounded strategy combines the strengths of supervised and unsupervised styles, creating a nuanced prediction system with high accuracy.

The research contributes to the healthcare system by reducing diabetic deaths, facilitating quick diagnosis of diabetes prediction, and improving the quality, precision, recall, accuracy, and f1-score of diabetic prediction models. The primary goal is to achieve desired results by proposing the use of PCA, Fuzzy C-Means, and Random Forest algorithms for diabetic prediction. Various ML techniques for diabetic prognosis are investigated and compared, with a focus on combining multiple machine learning methods. The proposed algorithm is evaluated against existing diabetic prediction

methods using well-known assessment criteria such as f1-score, recall, and accuracy.

2 LITERATURE REVIEW

Many scientists and doctors aim to identify diabetes in individuals as early as possible. In their collaborative research, computer algorithms and data mining were utilized to develop accurate, cost-effective, and rapid procedures for diabetes analysis. In the study by Iyer and his associates [1], the objective was to enhance the prediction of diabetic diagnosis using the Naive Bayes algorithm for categorizing diabetes diagnostic predictions. Positive results were achieved in 79.56 percent of cases.

In another significant paper by Tarun [2], it was reported that PCA and SVM algorithms were employed for classifying diabetic patients in the previously established diabetes diagnostic prediction model. The implementation success rate was found to be 93.66 percent. Other research papers included in a separate study report involved the development of an improved K-mean and logistic regression model for diabetes by Han et al. [3]. The model used logistic regression and the K-Means algorithm as implementation techniques, and the trial was 95.42 percent successful.

Gowda proposed [4] in another study to optimize the diabetic prognosis model using SVM and a clustering algorithm called fuzzy c-means, along with the execution technique. The data implementation success rate was 94.30 percent. In another paper, Patil predicted [5] that the classification forecasting model utilized C4.5 and the K-Means algorithm to enhance the hybrid prediction model and implementation technique, with findings showing 92.38 percent positive results.

Anjali suggested in a future study [6] that the examination technique would use PCA and the Neural Network (NN) algorithm to improve the diagnosis prediction classifier model. Executions were successful in 92.2 percent of cases. In a separate study report [7], Motka stated that the experimental approach used PCA and the Artificial Neural Fuzzy Interference System to retrieve the prior study diabetic prediction classifier model (ANFIS), with 89.2 percent of implementations being successful.

In another study, Chandigarh was intended [8] to develop a more accurate diabetes prediction model and implementation strategy based on PCA Data mining for descriptive purposes, Neural Network (NN), and Cultural Algorithm (CA). The execution results were 92.2 percent positive. Kumar predicted [9] in another study that the diabetic prediction classifier model and testing technique used the Support Vector Machine (SVM) algorithm, with a successful implementation rate of 78.0 percent.

According to Sanakal [10] in another paper, the goal was to enhance the diabetic prediction model's diagnosis, utilizing the SVM and Fuzzy C-means clustering method, resulting in a 94.30 percent success rate. Yilmaz proposed [11] using Support Vector Machine (SVM) and modified K-means algorithms to create a better diabetic prediction model, with executions successful in 96.71 percent of cases.

Changsheng, in another planned study [12], used the earliest and best diabetic prediction model and execution approach with PCA, K-Means, and Logistic Regression algorithm, achieving a 97.40 percent implementation success rate. The research then concluded with a comparative analysis of two prominent machine learning algorithms, k-Nearest Neighbor (KNN) and Naive Bayes, applied to predict diabetes using health attributes in the Pima Indians dataset.

The evaluation, conducted through a Confusion Matrix, revealed that the Naive Bayes algorithm exhibited better performance compared to KNN. The average evaluation metrics obtained for both algorithms were as follows:

Naive Bayes: 76.07% accuracy, 73.37% precision, and 71.37% recall. KNN: 73.33% accuracy, 70.25% precision, and 69.37% recall.

The findings strongly indicated that the Naive Bayes algorithm was more effective in predicting diabetes from the given dataset compared to KNN. This study suggested that Naive Bayes is preferred due to its high precision, accuracy, and recall [29].

The research concluded with a comparative analysis of two prominent machine learning algorithms, k-Nearest Neighbor (KNN) and Naive Bayes, applied to

predict diabetes using health attributes in the Pima Indians dataset. The findings strongly indicated that the Naive Bayes algorithm was more effective in predicting diabetes from the given dataset compared to KNN [30].

3 RESEARCH METHODOLOGY

The first step in this section is to provide an algorithm; data is loaded from a database, and then information is prepared. Following the completion of the data preparation process, the data is preprocessed, and standardized, and PCA is implemented. PCA works by first normalizing data in the database, then performing Principal Component Analysis and determining the result. After that, we formally classify the diagnostic prediction system with supervised classification using the Fuzzy C-Means Clustering algorithm for unsupervised clustering (because the Fuzzy C-Means algorithm cannot clean up and convert label data from the Fuzzy C Mean cluster result). That's because combining different methods resulted in more remarkable results. To develop the goal model, the benefits of PCA, Fuzzy C-Means, and the Random Forest Algorithm will be used. The testing data is 0.25 percent in size, and the implement train data is 0.75 percent in size. Figure 1 depicts a flowchart that may assist you in understanding the activity and briefly outlines the framework for dividing out work.

3.1 DATA COLLECTION

Data was obtained from the Kaggle website, which contains a variety of machine learning datasets, as well as the Pima Indians Diabetes dataset, which Changsheng had previously used [12]. The dataset contains 768 female patients from Arizona, United States, who underwent diabetes testing and had a total of nine attributes. (each attribute represents a measure of medical diagnosis) and one class called targeted class (that shows every individual test status), with 500 negative test cases and 268 positive cases in this set of data. Table 1 shows the dataset attribute explanation of Pima Indians Diabetes [12].

3.2 PCA (Principal Component Analysis)

PCA is a dimension reduction method that is widely used to deal with large amounts of data predictions by separating many variables into smaller ones and integrating much of the data from information machines into a massive data set. The number of items in the test set is limited, which undoubtedly reduces precision. The issue with dimensionality reduction is that it sacrifices rigidity for usability. Because smaller chunks of data are easier to examine and replicate, data evaluation is certainly less demanding and faster for machine learning models with free components for testing. PCA frequently employs preprocessing, depreciation measurement reduction, covariance and association eigenvalues and eigenvectors approaches. It also discusses data standardization, preprocessing, simulation and induction requirements, significance criteria, sophistication considerations, identified construct post-processing, simulation and instructional models, and data mining techniques. The key steps of the PCA algorithm are depicted in:

- The dataset needs to be normalized first.
- Determine the characteristics of the dataset's covariance matrix.
- Determine the eigenvalues and eigenvectors of the covariance matrix.
- Sort the eigenvalues and eigenvectors logically.
- Construct an eigenvector matrix of selected eigenvalues.
- Reconstruct the original matrix.

The PCA methodology can make use of the mathematical approaches and functions listed below. Standardization, covariance, Eigenvalues, and Eigenvectors can all be calculated using PCA [14], methods include the following:

Table 2 shows the dataset of Pima Indian Diabetes derived from the UCI machine-learning repository that was used in this study. The data set includes 768 female patients who were screened for diabetes in the Arizona population in the United States. The dataset includes eight properties that represent medical diagnostic criteria as well as a target class that represents each individual's status. The data set contains 268 positive cases and 500 negative cases. The dataset's attributes are as follows: 1) Expectant mothers, 2) Glucose, 3) Blood Pressure, 4) Skin Thickness, 5) Insulin

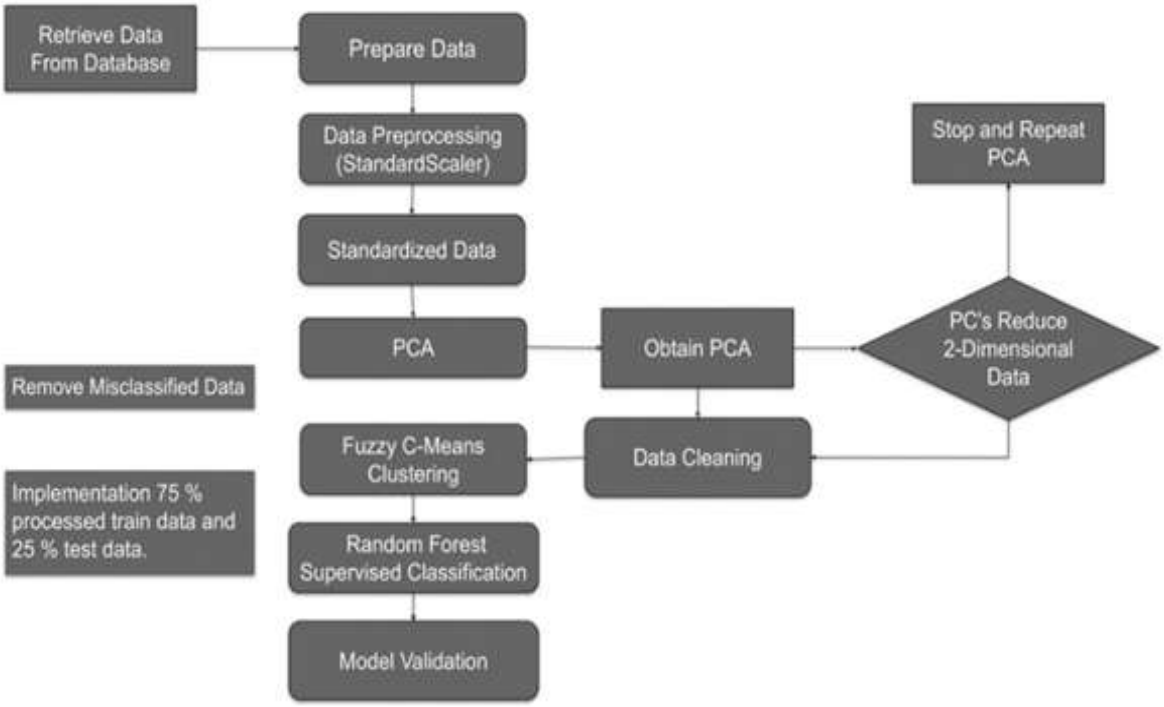

Figure 1. Flowchart for Goal Model Development Process

Table 1. Description of attributes in the Pima Indians Diabetes dataset.

S/No.	Attribute Name	Description
1	Pregnancies	How many times did pregnancy occur?
2	Glucose	In an oral glucose tolerance test, what is the plasma glucose concentration in two hours?
3	Blood Pressure	Diastolic and systolic blood pressure
4	Skin Thickness	Triceps skinfold thickness
5	Insulin	Two-hour serum insulin levels
6	BMI	The body mass index
7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	Age of the Patient
9	Outcome	The specific variable that is the focus or subject of analysis within the dataset.

Table 2. Basic Equations for the Model

Equation	Description
Standardization	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}}$
Covariance (for population)	$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$
Covariance (for sample)	$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{N-1}(x_i-\bar{x})(y_i-\bar{y})}{N-1}$
Eigenvalues and Eigenvectors	
$Av - \lambda v = 0; (A - \lambda I)v = 0$	A is the covariance matrix, λ is an eigenvalue, and v is an eigenvector.

Deficiency, 6) Body Mass Index (BMI), 7) Diabetes Pedigree Function, and 8) Age. Thirteen of the 44 pregnant women had diabetes by the end of the study. We can also conclude that older women are more likely to develop diabetes. Table 2 summarizes the features and variables used in the Pima Indians Diabetes implementation dataset for this study. Figure 2 depicts the dataset after standardization, with "x" and "y" values clearly defined.

The technique aided in lowering the downside of getting reproduction attributes which might be vain for grouping, which turned into a full-size result produced with the aid of using the usage of PCA. PCA progressed our Fuzzy C-Means outcomes due to the fact lowering the variety of things inside the actual set of information made it less difficult to cope with the doubtful or misclassified information. PCA gives a full-size gain with the aid of serving as an essential step in estimating the overall variety of clusters and offering a mathematical framework to specify the composition of those clusters. When the number one additives of the information were identified, we can use them to correctly compress the information. To accomplish this, the overall variety of dimensions ought to be decreased at the same time as maintaining a full-size quantity of applicable information. The effectiveness and consistency of diagnostic and predictive fashions are essential and have to be ensured earlier than use. To examine and examine our version output, we used lots of sets of rules assessment combinations. The wiped-clean and preprocessed information are proven in Table 3. Figure 3 depicts the 2-D information Principal Component Analysis (PCA) Data Chart in detail.

3.3 FUZZY C-MEANS CLUSTERING

Unsupervised learning consists of a type called clustering, which has numerous applications and standard implementations in various industries. Clustering is the separation and processing of data on behalf of an information system, resulting in data sets called clustering, and each cluster is assigned a primary ID. FCM algorithms are currently widely used. In this method, each data point is given a membership based on the separation between the cluster center and the data points corresponding to each cluster center. The closer the data is to the center of the cluster, the more it belongs to the center of the cluster. The membership for each data point must sum to 1 [15]. The formula shown below is used to change the membership and cluster center after each iteration.

Fuzzy C-Means equations:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}$$

$$v_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^m x_i}{\sum_{i=1}^{n}(\mu_{ij})^m}, \quad \forall j = 1, 2, 3, \ldots, c$$

The main objectives of the Fuzzy C-Means method are to reduce:

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m \|x_i - v_j\|^2$$

where $\|xi - vj\|$ is the Euclidean distance between the ith data point and the jth cluster center.

Table 3. Pima Indians Diabetes Dataset

P-ID	Pregnancies	Glucose	BP	S-Thickness	Insulin	BMI	DPF	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1

Table 4. Preprocessed Pima Indians Diabetes Dataset

Pregnancies	Glucose	BP	S-Thickness	Insulin	BMI	DPF	Age	Outcome
0.63994	0.848324	0.149641	0.907270	-0.692891	0.204013	0.468492	1.425995	1.365896
-0.84488	-1.123396	-0.160546	0.530902	-0.692891	-0.684422	-0.365061	-0.190672	-0.732120
1.23388	1.943724	-0.263941	-1.288212	-0.692891	-1.103255	0.604397	-0.105584	1.365896
-0.84488	-0.998208	-0.160546	0.154533	0.123302	-0.494043	-0.920763	-1.041549	-0.732120
-1.14185	0.504055	-1.504687	0.907270	0.765836	1.409746	5.484909	-0.020496	1.365896
0.34298	-0.153185	0.253036	-1.288212	-0.692891	-0.811341	-0.818079	-0.275760	-0.732120
-0.25095	-1.342476	-0.987710	0.719086	0.071204	-0.125977	-0.676133	-0.616111	1.365896
1.82781	-0.184482	-3.572597	-1.288212	-0.692891	0.419775	-1.020427	-0.360847	-0.732120
-0.54791	2.381884	0.046245	1.534551	4.021922	-0.189437	-0.947944	1.681259	1.365896
1.23388	0.128489	1.390387	-1.288212	-0.692891	-4.060474	-0.724455	1.766346	1.365896

Some distance methods:

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

$$\text{Manhattan:} \quad \sum_{i=1}^{k} |x_i - y_i|$$

$$\text{Minkowski:} \quad \left(\sum_{i=1}^{k} |x_i - y_i|^q\right)^{1/q}$$

FCM algorithm steps:

1. Let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ be a set of data points, and $V = \{v_1, v_2, v_3, \ldots, v_c\}$ be the center within the set.
2. Randomly select 'c' cluster centers.
3. Determine the membership of fuzzy μ_{ij} using the following formula:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}$$

We added the PCA dataset to this preprocessing to create a mixed data display. PCA is a technique used to reduce the size of these datasets, improve accessibility, and minimize data loss through Fuzzy C (two clusters). In a soft clustering technique known as FCM, each data point is given a probability or probability score that indicates whether it belongs to that group. This is done by creating new uncorrelated variables that continuously maximize their variance. This logic was quickly implemented in the data matrix to create a grouping matrix showing the

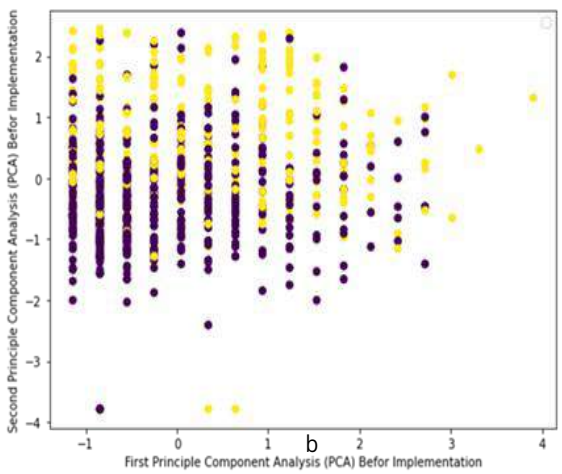

Figure 2. Dataset after standardization

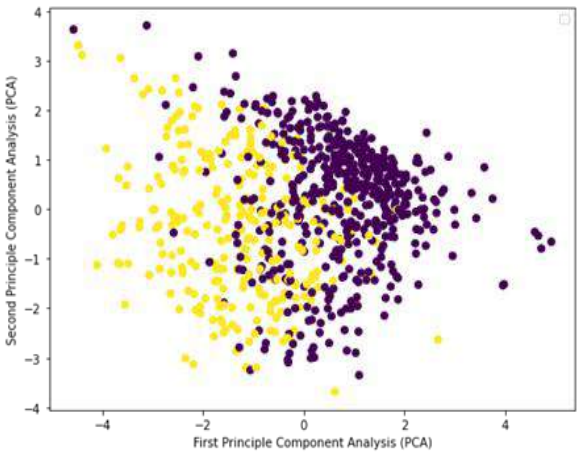

Figure 3. Principal Component Analysis (PCA)

relationship between each cluster and sample.

3.4 RANDOM FOREST ALGORITHM

Random forest is a machine-learning technique used to resolve regression and classification conflicts. It makes use of ensemble methods, which are a type of problem-solving system that makes use of categorization techniques. It combines many classifiers to tackle a difficult task and improve the model's effectiveness. The RF method, which is based on decision tree prediction, determines the outcome. It makes predictions by averaging or computing the results of multiple trees. The results show that as the number of trees increases, the precision improves. The constraints of a decision tree algorithm are removed using an RF. The Random Forest approach reduces dataset overfitting and improves precision, recall, and f1-score, among other things, for the reasons listed below [16][17].

- In comparison to other algorithms, it takes less time.
- It forecasts output with high precision and efficiency, particularly when dealing with large datasets.
- It can maintain its accuracy even when a significant portion of the data is missing.

The random forest procedure is based on the following fundamental principles:

- Begin by randomly selecting samples from a given dataset.
- The algorithm then generates a decision tree for each and every sample. The expected outcomes for each decision tree will then be calculated. Each predicted outcome will be nominated or voted on.
- Finally, as the most accurate prediction, choose the most popular forecast result.

Figure 4 shows how the Random Forest Algorithm works [18].

The Random Forest Algorithm technique, on the other hand, employs some analytical functions or formulas. This technique, as well as Gini (Coefficient, Index, or Ratio), Entropy, and Mean Squared Error (MSE) [19], can be used by the Random Forest Algorithm.

4 RESULTS AND DISCUSSION

To obtain accurate results, we compare several accuracy models of Machine Learning and Deep

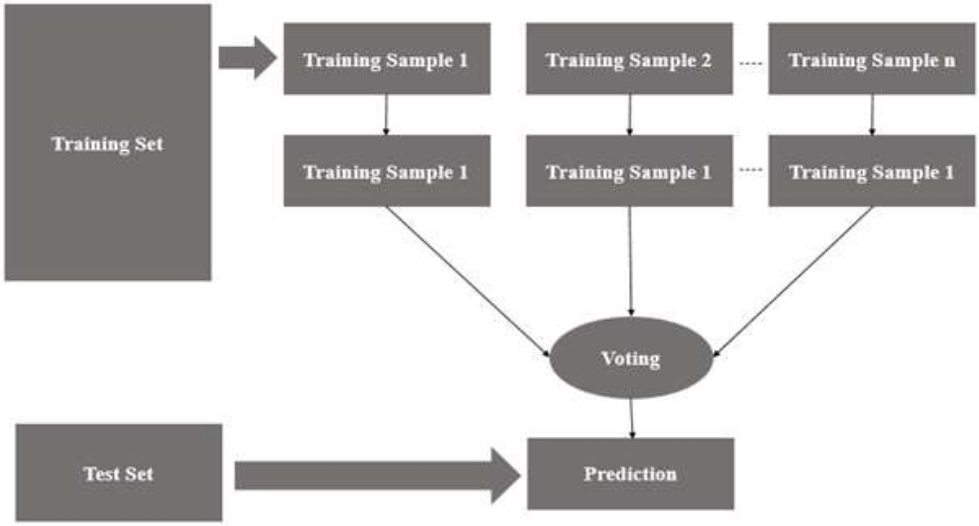

Figure 4. Mechanism of Random Forest Algorithm

Learning in this study to see which fits best to our dataset and provides the most accuracy. Machine Learning refers to computers learning from data and using algorithms to perform a task that has not been explicitly programmed. Deep Learning, on the other hand, employs a complex structure of algorithms modeled after the human brain. This allows unstructured data such as documents, images, and text to be processed. On specific datasets, we use both Machine Learning and Deep Learning Algorithms in our research. The highest accuracy was obtained when combining PCA with Fuzzy-C-Means and Random Forest in the Algorithms Model Accuracy Diabetes Diagnosis Prediction - Table 4.

To obtain the precise Precision, Recall, and F1-Score of our result in Table 5, we used the same accuracy models as in the previous table to determine which best fit our dataset. We have listed all outcomes in the combination of Algorithms Model Accuracy Diabetes Diagnosis Prediction.

We had to look over the results of various accuracy models as we processed them. To name a few, PCA with Fuzzy-C-Means and XGBoost, Fuzzy-C-Means and K-Nearest Neighbors, and PCA Fuzzy-C-Means and Decision Tree. Furthermore, precision, recall, and F1 score are not going away. Finally, Figure 5 shows how the ROC curve balances specificity (or false positive rate) and sensitivity (positive rate). Classifiers that have curves that are closer to the upper left corner perform better. When the curve approaches the ROC space's 45-degree diagonal line, the test's accuracy decreases.

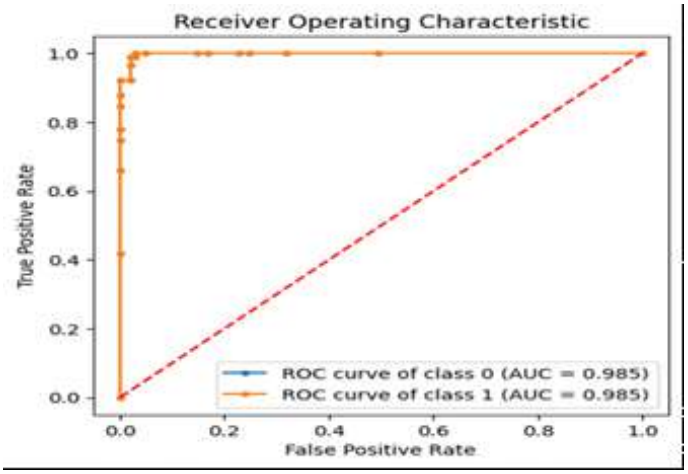

Figure 5. ROC curve

Figure 5: shows how the ROC curve balances specificity (or false positive rate) and sensitivity

Table 5. Combination of Algorithms Model Accuracy for Diabetes Diagnosis Prediction

S/No.	Combination of Algorithms	Accuracy (%)
1	PCA, Fuzzy-C-Means, Naïve Bayes	94.7916
2	PCA, Fuzzy-C-Means, Decision Tree	96.3541
3	Fuzzy-C-Means, Naïve Bayes	96.875
4	PCA, Fuzzy-C-Means, XGBoost	96.875
5	PCA, Fuzzy-C-Means, K-Nearest Neighbors	97.3958
6	Fuzzy-C-Means, K-Nearest Neighbors	97.3958
7	Fuzzy-C-Means, Logistic Regression	97.9166
8	PCA, Fuzzy-C-Means, Simple Recurrent Neural Networks	97.92
9	PCA, Fuzzy-C-Means, Long Short-Term Memory	97.92
10	Fuzzy-C-Means, Random Forest	98.4375
11	Fuzzy-C-Means, Support Vector Machine	98.4375
12	Fuzzy-C-Means, XGBoost	98.4375
13	Fuzzy-C-Means, Decision Tree	98.4375
14	Proposed Method (PCA, Fuzzy-C-Means, Random Forest)	98.9782

(positive rate) Classifiers that have curves that are closer to the upper left corner perform better. When the curve approaches the ROC space's 45-degree diagonal line, the test's accuracy decreases.

5 CONCLUSION

The goal of this study was to develop an accurate model for diabetes prediction. We presented a new model after a thorough review of other published papers that included PCA to reduce dimensionality, Fuzzy-C-Means for classification, and Random Forest. We first used the PCA technique on our data set to improve the k-means results of other researchers. Although PCA is a well-known technique, its effectiveness in improving Fuzzy-C-Means and the RF model of classification has received little attention. In our experiment, we demonstrated that combining PCA and Fuzzy-C-Means improves the accuracy of a Random Forest model for diabetes prediction. The study's uniqueness is its ability to produce an improved Fuzzy-C-Means classifier result that is far superior to what other researchers have achieved in similar studies. In contrast to the results obtained using other algorithms in our study, the Random Forest model improved the prediction of diabetes onset. Another feature is that our model can run successfully on new datasets. The researchers improved the diabetic prediction model by using PCA, the Logistic Regression algorithm, and K-Means, achieving an accuracy level of 97.40%. Using PCA with Fuzzy-C-Means and Random Forest, the accuracy level was increased to 98.96Potential Conflict of Interest The authors had no conflict of interest.

Acknowledgment

I sincerely thank my Ph.D. advisors, Dr. Abdul Qadar Kara, Dr. Imran Jami, Dr. Naseem, and Dr. Saood Zia, for their invaluable guidance. Special gratitude to our dedicated research team whose contributions were crucial to the project's success.

Author Contributions

Priha Bhatti: Conceptualization, Methodology, Data curation, Writing- Original draft preparation , Visualization, Investigation. **Iqra Heer Bhatti: Noorulain Kamran**: Investigation **Khalid Mahboob, Saad Naeem**: Software, Validation.

Table 6. Algorithm Combination Precision, Recall, and F1-Score for Diabetes Diagnosis Prediction

S/No.	Combination of Algorithms	Precision (%)	Recall (%)	F1-Score (%)
1	PCA, Fuzzy-C-Means, Naïve Bayes	93.54	95.60	94.56
2	PCA, Fuzzy-C-Means, Decision Tree	94.68	97.80	96.21
3	Fuzzy-C-Means, Naïve Bayes	98.55	93.15	97.29
4	PCA, Fuzzy-C-Means, XGBoost	95.69	97.80	96.73
5	PCA, Fuzzy-C-Means, K-Nearest Neighbors	96.73	97.80	97.26
6	Fuzzy-C-Means, K-Nearest Neighbors	98.57	98.57	96.50
7	Fuzzy-C-Means, Logistic Regression	98.59	95.89	97.22
8	PCA, Fuzzy-C-Means, Simple Recurrent Neural Networks	97.98	97.85	97.90
9	PCA, Fuzzy-C-Means, Long Short-Term Memory	97.91	97.91	97.91
10	Fuzzy-C-Means, Random Forest	98.61	97.26	97.93
11	Fuzzy-C-Means, Support Vector Machine	98.61	97.26	97.93
12	Fuzzy-C-Means, XGBoost	98.61	97.26	97.93
13	Fuzzy-C-Means, Decision Tree	98.61	97.26	97.93
14	Proposed Method (PCA, Fuzzy-C-Means, Random Forest)	97.84	100.0	98.56

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

Funding Information

No Funding received for the research.

References

[1] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," International Journal of Data Mining and Knowledge Management Process (IJDKP), vol. 5, no. 1, 2015.

[2] T. Jhaldiyal and P. K. Mishra, "Analysis and prediction of diabetes mellitus using PCA, REP and SVM," International Journal of Engineering and Technology Research (IJETR), vol. 2, issue 8, ISSN: 2321-0869, 2014.

[3] W. Han, S. Y. Shengqi, H. Zhangqin, 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.

[4] G. K. Asha, 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 Applications, vol. 45, no. 12, ISSN: 0975 – 8887, 2012.

[5] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," Expert Systems with Applications, vol. 37, pp. 8102–8108, 2010.

[6] A. Khandegar and K. Pawar, "Diagnosis of Diabetes Mellitus Using PCA, Neural Network and Cultural Algorithm," International Journal of Digital Application Contemporary Research, vol. 6, ISSN: 2319-4863, 2017.

[7] M. Rakesh, P. Viral, K. Balbindra, and A. R. Verma, "Diabetes mellitus forecast using different data mining techniques," Proceedings of the IEEE 4th International Conference on Computer and Communication Technology (ICCCT), pp. 99–103, IEEE, 2013.

[8] A. Khandegar, "Diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm," International Journal of Digital Application Contemporary Research, vol. 5, no. 6, 2017.

[9] A. Kumari and R. Chitra, "Classification of Diabetes Disease Using Support Vector Machine," International Journal of Engineering Research and

Applications (IJERA), March-April, pp. 1797-1801, ISSN: 2248-9622, 2013.

[10] S. Sanakal and S. T. Jayakumari, "Prognosis of diabetes using data mining approach - Fuzzy C means clustering and support vector machine," International Journal of Computer Trends and Technology (IJCTT), vol. 11, no. 2, 2014.

[11] N. Yilmaz, O. Inan, and M. S. Uzer, "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," Journal of Medical Systems, vol. 38, no. 5, 2014.

[12] C. Zhu, C. U. Idemudiaa, and W. Fengb, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Journal of Medical Imaging, ISSN: 2352-9148, 2019.

[13] Diabetes Daily, [Online]. Available: https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/.

[14] Medium - Understanding Principal Component Analysis (PCA) Step by Step, [Online]. Available: https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9.

[15] Data Clustering Algorithms - Fuzzy C-Means Clustering Algorithm, [Online]. Available: Data Clustering Algorithms

[16] Section.io - Introduction to Random Forest in Machine Learning, [Online]. Available: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

[17] Tutorials Point - Machine Learning with Python: Random Forest Classification Algorithms, [Online]. Available: https://www.tutorialspoint.com

[18] Tutorials Point - Machine Learning with Python: Random Forest Algorithm Image, [Online]. Available: https://www.tutorialspoint.com/machine learning with python.

[19] R. Huss, J. Raffler, and B. Märkl, "Artificial intelligence and digital biomarker in precision pathology guiding immune therapy selection and precision oncology," Cancer Reports, e1796, 2023.

[20] S. Karim, A. Qadir, U. Farooq, M. Shakir, and A. Laghari, "Hyperspectral imaging: a review and trends towards medical imaging," Current Medical Imaging, vol. 19, no. 5, pp. 417-427, 2023.

[21] A. V. Singh, V. Chandrasekar, N. Paudel, P. Laux, A. Luch, D. Gemmati, V. Tissato, K. S. Prabhu, S. Uddin, and S. P. Dakua, "Integrative toxicogenomics: Advancing precision medicine and toxicology through artificial intelligence and OMICs technology," Biomedicine Pharmacotherapy, vol. 163, 114784, 2023.

[22] B. Ndzendze and T. Marwala, "Artificial Intelligence and International Relations," Springer Nature Singapore, pp. 33-54, 2023.

[23] A. A. Khan, A. A. Laghari, and S. A. Awan, "Machine learning in computer vision: a review," EAI Endorsed Transactions on Scalable Information Systems, vol. 8, no. 32, e4-e4, 2021.

[24] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer," Neuroscience Informatics, vol. 2, no. 4, 100034, 2022.

[25] D. K. K. Reddy, H. S. Behera, J. Nayak, A. R. Routray, P. S. Kumar, and U. Ghosh, "A Fog-Based Intelligent Secured IoMT Framework for Early Diabetes Prediction," in Intelligent Internet of Things for Healthcare and Industry, Springer International Publishing, pp. 199-218, 2022.

[26] P. M. Lozano, M. Lane-Fall, P. D. Franklin, R. L. Rothman, R. Gonzales, M. K. Ong, M. K. Gould, et al., "Training the next generation of learning health system scientists," Learning Health Systems, vol. 6, no. 4, e10342, 2022.

[27] A. A. Laghari and S. Yin, "How to Collect and Interpret Medical Pictures Captured in Highly Challenging Environments that Range from Nanoscale to Hyperspectral Imaging," Current Medical Imaging, 2022.

[28] R. Chauhan, A. Goel, H. Kaur, and B. Alankar, "Machine Learning: An Analytical Approach for Pattern Detection in Diabetes," in Soft Computing: Theories and Applications: Proceedings of SoCTA 2022, Springer Nature Singapore, pp. 135-145, 2022.

[29] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Computer Science, vol. 216, pp. 21-30, 2023, DOI: 10.1016/j.procs.2022.12.107.

[30] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Computer Science, vol. 7, issue 4, pp. 432-439, December 2021, DOI: 10.1016/j.procs.2022.12.107.

[31] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki, P. K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," J. Healthc. Eng., vol. 2022, Art. no. 1684017, 2022. [Online]. Available: https://doi.org/10.1155/2022/1684017

[32] U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.