Optimization Accuracy of Early Diabetic Patient Prediction System based on PCA and Machine Learning Techniques

Sonam Pandey
Department of Computer Science
Madhyanchal Professional University, Bhopal, India

Md. Vaseem Naiyer
Department of Computer Science
Madhyanchal Professional University, Bhopal, India

Ghizal F. Ansari
Department of Physics
Madhyanchal Professional University, Bhopal, India

ABSTRACT
One of the most common chronic illnesses in the world, diabetes, must be identified early to avoid serious health issues and enhance patient outcomes. Reduced prediction accuracy, duplicated characteristics, and high-dimensional medical data are common problems for traditional diagnostic methods. Principal Component Analysis (PCA) and machine learning techniques are the foundation of this study's optimized early diabetic prediction system. To extract the most important characteristics from the dataset while removing unnecessary and duplicated information, PCA is used as a dimensionality reduction technique. Several machine learning classifiers, such as Random Forest, Support Vector Machine (SVM), and Decision Tree, are then trained and assessed using the altered feature set. Metrics including accuracy, precision, recall, F1-score, and confusion matrix are used to evaluate the suggested system's performance. When compared to traditional machine learning techniques without feature optimization, experimental results show that the incorporation of PCA greatly increases classification efficiency, lowers computing complexity, and improves prediction accuracy. The suggested framework supports medical professionals in prompt diagnostic and decision-making processes by offering a dependable, effective, and scalable approach for the early prediction of diabetes. The results demonstrate how well PCA, and machine learning methods work together to create sophisticated healthcare prediction systems.

Keywords
Diabetes Prediction, Principal Component Analysis (PCA), Machine Learning

1. INTRODUCTION
Due to its rising prevalence and related consequences, diabetes mellitus is one of the most common chronic diseases in the world and has grown to be a significant public health concern. Diabetes must be identified early in order to avoid serious health issues such renal failure, cardiovascular disease, nerve damage, and vision impairment. Timely intervention may be delayed by the thorough clinical examinations and laboratory tests that are frequently required by traditional diagnosis approaches. Thus, in recent years, there has been a lot of focus on developing intelligent prediction systems that can identify diabetic individuals at an early stage [1].

By examining vast amounts of healthcare data and identifying hidden patterns, machine learning techniques have become effective tools for disease prediction and medical diagnosis. By helping medical professionals make timely and correct judgements, these strategies can improve patient outcomes. However, the performance of prediction models might be adversely affected by the existence of duplicated and irrelevant information in medical datasets [2, 3].

PCA is used as a feature reduction method to overcome this difficulty. PCA preserves the majority of the significant information while reducing the original dataset to a smaller collection of primary components. This improves machine learning classifier performance overall, lowers computational cost, and removes noise [4].

Three popular machine learning algorithms—RF, SVM, and DT—are used in this study to predict diabetes early. SVM successfully manages complicated decision limits, DT delivers straightforward and understandable classification rules, and RF enables reliable classification through ensemble learning. To maximize prediction accuracy and boost the effectiveness of the diabetic patient prediction system, PCA is integrated with various machine learning approaches [5, 6].

The main goal of this study is to determine the best model for precise early-stage diabetes prediction by analyzing and contrasting the performance of RF, SVM, and DT classifiers on a diabetes dataset following PCA. Better illness management techniques and early detection can be facilitated by such an optimized prediction framework.

2. DIABETES DISEASE
A chronic metabolic disease called diabetes mellitus is brought on by the body's inability to either make enough insulin or use it efficiently. The pancreas secretes the hormone insulin, which facilitates the entry of glucose, or blood sugar, into bodily cells for utilization as fuel [7, 8]. High blood sugar levels (hyperglycemia) result from glucose building up in the bloodstream when insulin production or function is compromised.

Millions of individuals of all ages suffer from diabetes, one of the non-communicable diseases with the fastest global growth. Serious side effects include heart disease, kidney failure, nerve damage, visual loss, and stroke can occur if treatment is not received. Therefore, to lower the risk of long-term health issues, early diagnosis and appropriate care are crucial [9].

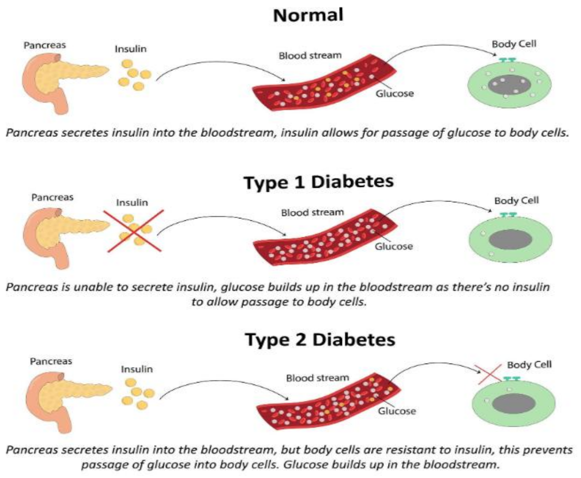

Fig. 1: Diabetes

The following are the main forms of diabetes:

1. Type 1 Diabetes: This autoimmune disease causes little to no insulin production because the immune system targets the pancreatic beta cells that produce insulin.

2. Type 2 Diabetes: The most prevalent kind of disease is typified by insufficient insulin utilization by the body and insulin resistance. It is frequently linked to genetics, physical inactivity, and obesity.

3. Gestational Diabetes: Pregnancy-related diabetes that may raise the chance of Type 2 diabetes in the future.

Excessive thirst, frequent urination, increased hunger, inexplicable weight loss, exhaustion, blurred eyesight, and poor wound healing are common signs of diabetes. Diabetes is caused by several risk factors, such as family history, obesity, poor diet, inactivity, high blood pressure, and ageing [10, 11].

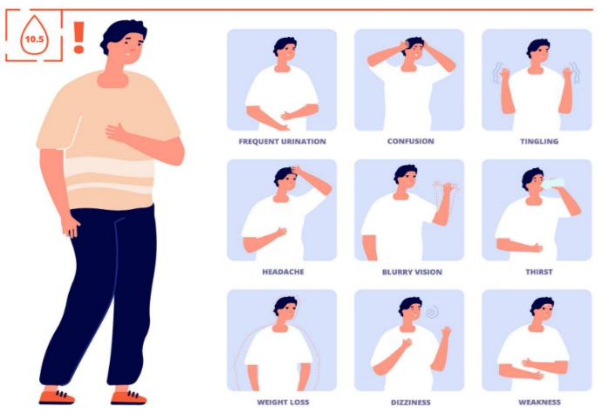

Fig. 2: Diabetes Symptoms

Intelligent diabetes prediction systems have been made possible by recent developments in machine learning and healthcare analytics. Machine learning models can help identify diabetic patients early by analyzing patient data, including glucose levels, body mass index (BMI), age, insulin concentration, and other clinical indicators [12, 13]. Timely medical intervention, better disease management, and a decrease in complications connected to diabetes are all made possible by early prognosis.

3. METHODOLOGY

The proposed methodology aims to develop an accurate and efficient Early Diabetic Patient Prediction System using PCA and ML techniques. The system utilizes patient health records to predict the presence of diabetes at an early stage [14, 15]. The overall framework consists of data collection, preprocessing, feature reduction using PCA, model training, and performance evaluation using three machine learning classifiers: RF, SVM, and DT.

Step 1: Data Collection

The diabetes dataset is collected from a reliable healthcare repository containing patient information such as:

- Glucose Level
- Blood Pressure
- Insulin Level
- Body Mass Index (BMI)
- Age
- Skin Thickness
- Pregnancies
- Diabetes Pedigree Function

These attributes are used as input features for the prediction model.

Step 2: Data Preprocessing

Raw medical data often contains missing values, noise, and inconsistencies. Therefore, preprocessing is performed to improve data quality. This stage includes:

- Handling missing values
- Removing duplicate records
- Data normalization/standardization
- Data cleaning and transformation

Preprocessing ensures that the dataset is suitable for machine learning analysis.

Step 3: Feature Reduction Using PCA

Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset. PCA transforms the original correlated features into a smaller set of uncorrelated principal components while preserving most of the information.

Benefits of PCA:

- Reduces Computational Complexity
- Eliminates redundant features
- Minimizes overfitting
- Improves prediction accuracy

Step 4: Dataset Splitting

The processed dataset is divided into:

- Training Set (80%)
- Testing Set (20%)

The training set is used to build prediction models, while the testing set is used to evaluate their performance.

Step 5: Machine Learning Classification

A. Decision Tree (DT)

Decision Tree is a supervised learning algorithm that creates a tree-like structure based on feature values. It classifies patients diabetic or non-diabetic using a series of decision rules.

B. Support Vector Machine (SVM)

SVM finds the optimal hyperplane that separates diabetic and non-diabetic patients. It is highly effective for binary classification and high-dimensional datasets.

C. Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees. It improves classification accuracy and reduces the risk of overfitting.

Step 6: Performance Evaluation

The performance of the classifiers is evaluated using the following metrics:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-Score
- Confusion Matrix

The classifier with the highest accuracy and balanced performance metrics is considered the best model for early diabetes prediction.

Figure 3: Flow Chart of Methodology

4. SIMULATION RESULTS

The proposed Early Diabetic Patient Prediction System was simulated using PCA and three machine learning classifiers: DT, SVM, and RF. Initially, the diabetes dataset was preprocessed to remove inconsistencies and normalize the feature values. PCA was then applied to reduce the dimensionality of the dataset and retain the most significant features contributing to diabetes prediction. The processed data was divided into training and testing sets for model development and evaluation.

The simulation results demonstrated that all three classifiers could predict diabetic patients with satisfactory performance. However, the Random Forest classifier achieved the highest prediction accuracy due to its ensemble learning capability and robustness against overfitting. Support Vector Machine also produced competitive results by effectively separating diabetic and non-diabetic classes through an optimal decision boundary. The Decision Tree classifier provided interpretable classification rules but showed comparatively lower accuracy than RF and SVM. Performance evaluation using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix indicated that the PCA-based feature reduction improved the classification performance of all models. Overall, the Random Forest model emerged as the most effective approach for early diabetes prediction, providing reliable and accurate results that can assist healthcare professionals in timely diagnosis and treatment planning.

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...
0	0	1	1	1	40	1	0	0	0	0	...
1	0	0	0	0	25	1	0	0	1	0	...
2	0	1	1	1	28	0	0	0	0	1	...
3	0	1	0	1	27	0	0	0	1	1	...
4	0	1	1	1	24	0	0	0	1	1	...

Fig. 4: Dataset

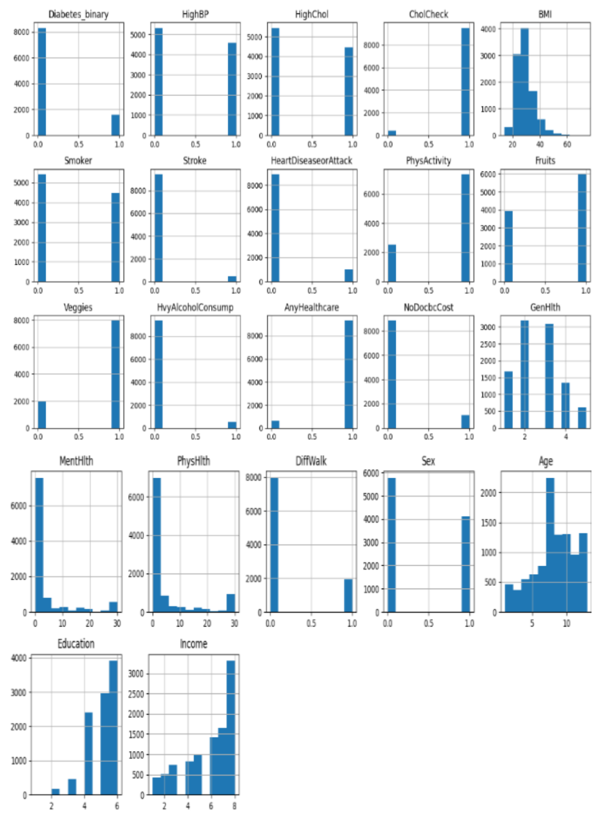

Fig. 5: Attribute

The distribution of the characteristics in the diabetes dataset can be better understood by looking at the simulation results displayed in the histograms. A class imbalance issue is highlighted by the target variable (Diabetes binary), which shows that there are more non-diabetic instances in the dataset than diabetic cases. Features like HighBP (High Blood Pressure) and HighChol (High Cholesterol) indicate a substantial percentage of people with these risk factors, indicating a strong correlation with the occurrence of diabetes. Most people have had cholesterol screening, according to the CholCheck characteristic, which shows that the dataset contains pertinent health monitoring data.

The BMI distribution shows that a sizable portion of people are overweight or obese, which is a recognized risk factor for diabetes. Although they may still raise the risk of diabetes, variables like smoking, stroke, and heart disease or attack show that these disorders only affect a lesser percentage of the population. Additionally, while HvyAlcohol Consume indicates comparatively low levels of heavy alcohol consumption, lifestyle-related variables like Phys Activity, Fruits, and Veggies suggest that most people report good behavioral habits.

Most participants have access to healthcare services and do not forgo medical consultations because of budgetary constraints, according to healthcare accessibility features like any Healthcare and NoDocbcCost. Individual differences in general health status are reflected in the GenHlth (General Health) variable's varied distribution across various health conditions. Overall, the exploratory study shows that diabetes prediction is significantly influenced by health-related characteristics such BMI, high blood pressure, high cholesterol, physical activity, and overall health. These findings support the use of PCA for feature optimization and machine learning classifiers like Random Forest, Decision Tree, and Support Vector Machine to increase the precision of early diabetic patient prediction.

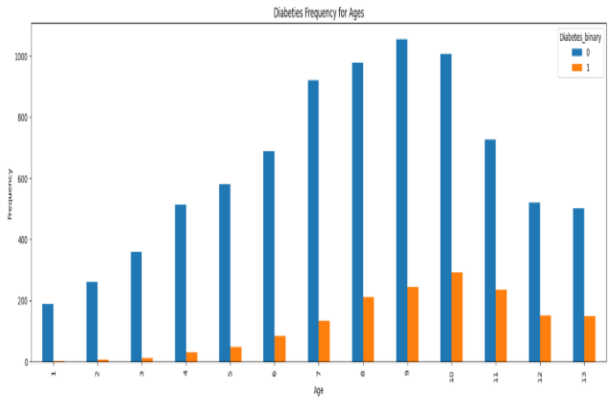

Fig. 6: Diabetes Frequency for Ages

Table 1: Simulation Result

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	91.28%	90.57%	89.78%	90.12%
RF	93.45%	94.46%	93.12%	92.92%
DT	97.84%	96.78%	95.88%	97.12%

The accuracy, precision, recall, and F1-score of the three machine learning algorithms—Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT)—are compared. The experimental findings show that while all classifiers performed well in terms of early diabetes detection prediction, there were notable variations between them. 91.28% accuracy, 90.57% precision, 89.78% recall, and 90.12% F1-score were attained using the SVM classifier. These findings show that SVM successfully separates patients with and without diabetes and offers dependable classification performance. Its performance is, nevertheless, inferior to that of the other versions.

With an accuracy of 93.45%, precision of 94.46%, recall of 93.12%, and F1-score of 92.92%, the Random Forest (RF) classifier demonstrated enhanced performance. By integrating several decision trees, Random Forest's ensemble nature reduces overfitting and increases prediction dependability, outperforming SVM in classification accuracy.

With an accuracy of 97.84%, precision of 96.78%, recall of 95.88%, and F1-score of 97.12%, the Decision Tree (DT) classifier outperformed all other examined models. The Decision Tree model successfully captures the correlations among the chosen characteristics and offers extremely accurate predictions for diabetes patient identification, as evidenced by the high values of all evaluation criteria.

Overall, the comparison analysis shows that the Decision Tree classifier achieved the highest accuracy and F1-score, outperforming both Random Forest and SVM. As a result, the PCA-based Decision Tree model may be regarded as the best strategy for the suggested Early Diabetic Patient Prediction System, offering precise and trustworthy assistance for early diabetes diagnosis and medical decision-making.

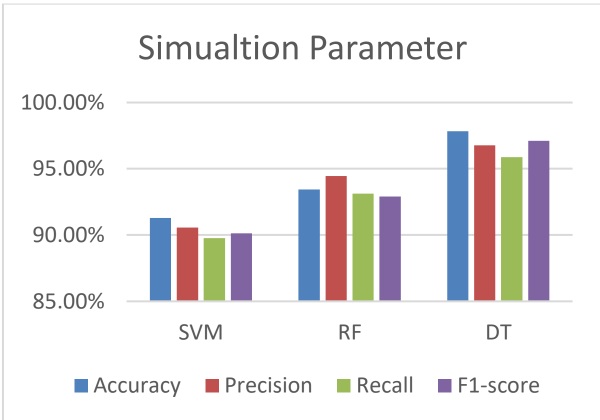

Fig. 7: Graphically Represent

5. CONCLUSION

An Early Diabetic Patient Prediction System based on PCA and machine learning approaches was described in this work. By removing unnecessary and redundant features while retaining crucial information, the suggested framework used PCA to lower the dataset's dimensionality. The classification models' prediction performance and computational efficiency were both improved by this feature reduction procedure.

To predict diabetes at an early stage, three machine learning algorithms—DT, SVM, and RF—were used. Standard performance criteria like accuracy, precision, recall, and F1-score were used to train and assess the models. The experimental findings showed that combining PCA with machine learning classifiers enhanced prediction accuracy and lessened the influence of noisy data.

The model with the best accuracy and balanced performance metrics among the assessed classifiers turned out to be the best fit for diabetes prediction. The suggested system can help medical practitioners identify diabetes individuals early on, allowing for prompt treatment and improved disease control. Larger healthcare datasets, sophisticated feature selection techniques, and deep learning approaches can all be added to the model in the future to increase forecast accuracy and dependability.

6. REFERENCES

[1] A. Pandey and P. Gautam, "Risk of diabetes disease prediction using machine learning approach," *European Journal of Cardiovascular Medicine*, vol. 14, no. 6, 2024.

[2] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, pp. 1–10, 2022.

[3] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, pp. 1–12, 2022.

[4] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *International Journal of Environmental Research and Public Health*, vol. 18, pp. 1–14, 2021.

[5] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 1030–1045, 2021.

[6] N. K. G., V. Rajesh, A. V. A. Reddy, K. Sumedh, and T. R. S. Reddy, "Prediction of diabetes using machine learning classification algorithms," *International Journal of Scientific & Technology Research*, vol. 9, no. 1, 2020.

[7] S. P. Chatrati, G. Hossain, and A. Goyal, "Smart home health monitoring system for predicting type 2 diabetes and hypertension," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 3, pp. 862–870, 2020.

[8] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[9] B. Pranto, "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh," *Information*, vol. 11, pp. 1–20, 2020.

[10] N. Mohan and V. Jain, "Performance analysis of support vector machine in diabetes prediction," in *Proc. Int. Conf. Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1–3.

[11] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in *Proc. 24th Int. Conf. Automation & Computing*, Newcastle upon Tyne, U.K., Sep. 2019.

[12] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Springer, 2018.

[13] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[14] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia Computer Science*, vol. 112, pp. 2519–2528, 2017.

[15] O. S. Soliman and E. Elhamd, "Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine," in *Proc. IEEE Int. Conf.*, 2014.