International Journal of Transdisciplinary Research and Perspectives
Volume 2, Issue 4, April 2026
www.ijtrp.com | E-ISSN: 3107-7935 | RESAIM Publishing | www.resaim.com

Diabetes Early Diagnosis System Using Machine Learning and Prognostic Feature Analysis

Shashwat Rai[1*], Gaurav Singh[1], Sunil Kumar[1]
[1]*Department of Computer Science and Engineering, Galgotias University, Greater Noida, India*

***Abstract*: Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycemia resulting from insulin deficiency or insulin resistance, and it has become a major global health concern due to its rapidly increasing prevalence. Since diabetes often remains asymptomatic in its early stages, late diagnosis leads to increased morbidity, mortality, and healthcare burden, making early detection essential. Recent advances in machine learning have enabled effective early disease diagnosis by analyzing large-scale medical data and identifying complex patterns beyond traditional methods. This paper presents a machine learning-based Diabetes Early Diagnosis System (DEDS) that utilizes prognostic feature analysis. The proposed system integrates data preprocessing, dimensionality reduction, feature selection, and multiple classification models to achieve accurate early-stage diabetes prediction. The PIMA Indian Diabetes Dataset is used for evaluation, incorporating clinical features such as glucose concentration, blood pressure, insulin level, body mass index, age, and hereditary factors. Principal Component Analysis (PCA) and Sparse PCA are applied to eliminate redundant features, improving model generalization and interpretability. Experimental results demonstrate that the proposed system outperforms traditional classifiers in terms of accuracy, sensitivity, and specificity, highlighting its effectiveness for early diabetes detection and its potential application in real-world clinical and telemedicine-based healthcare systems.**

***Keywords*: Diabetes Mellitus, Early Diagnosis, Machine Learning, Feature Selection, Prognostic Analysis, Healthcare Analytics.**

1. Introduction

Diabetes mellitus is a significant public health issue that is among the rapidly increasing non-communicable diseases in the world. It is defined by persistent hyperglycemia due to insulin deficiency or insulin resistance and is characterized by the extreme complication of the cardiovascular, kidney, eye, and nervous systems [3]. The increased incidence of diabetes has posed a substantial burden to the healthcare expenses and diminished the quality of life of a patient, so early detection of the condition is crucial to avoid complications and enhance clinical outcomes [1].

The classical diagnostic tools used in diagnosing diabetes, which include fasting plasma glucose, oral glucose tolerance test, and glycated haemoglobin (HbAlc), are clinically sound yet may necessitate numerous clinical appointments, special facilities and qualified staff [22]. Such restrictions can slow the diagnosis process especially in low-resource environments and might not be effective in detecting pre-diabetic states where the intervention can be most helpful. As such the demand on automated and cost-effective early diagnosis systems is increasing.

The acceleration in the development of electronic health records and medical databases has made it possible to apply machine learning methods in healthcare analytics. Models of machine learning are capable of processing large-scale medical data, detecting more complex patterns that can hardly be observed due to using traditional statistical methods [14]. Other researchers have already proved that a classifier like Logistic Regression, Support Vector Machines, Decision Trees, random forests, or Artificial Neural networks can effectively predict diabetes [6], [7], [13]. integrated into telemedicine and smart healthcare platforms. Nevertheless, most of the current methods are vulnerable to the issues of missing values, noisy data, class imbalance, and high dimensionality, which may adversely affect model performance and generalization [11], [22]. Moreover, there is a tendency to provide a low prioritization to the prognostic feature analysis, which reduces the clinical interpretability. In order to address such shortcomings, the paper suggests a Diabetes Early Diagnosis System (DEDS) that combines machine learning with the analysis of prognostic features. The system uses data preprocessing, dimensionality reduction by the PCA, Sparse PCA and various classification models to enhance prediction accuracy and performance. With the help of benchmark information provided by the PIMA Indian Diabetes Dataset [2], the suggested strategy should facilitate clinical decision making and allow efficient implementation in real-world and telemedicine-driven health care settings [15], [21].

2. Related Work

Machine learning has had comprehensive applications in diabetes prediction because it has the capability to identify intricate medical data and enhance diagnostic accuracy. Kavakiotis et al. [1] gave an extensive overview of machine

*Corresponding author: saispc0706@gmail.com

learning and data mining methods in diabetes studies, the focus being the use in clinical decision support systems. Logistic Regression is known to be easy and interpretable, however, it has difficulty in capturing non-linear relationships in medical data [13].

Support Vector Machines (SVMs) have demonstrated good performance in classifications by maximizing separation of margins yet they are sensitive to the choice of kernel and parameter optimization [7]. The widespread use of Decision Tree-based models, especially the Random Forests, is explained by the possibility to deal with nonlinear data, cope with the values missing, and estimate the feature importance measure that leads to better generalization [8]. The generally best methods to used in diabetes prediction tasks are ensemble learning methods, which are better than using single classifiers.

The use of deep learning techniques such as Artificial Neural Networks has also been used to diagnose diabetes and they have been highly predictive and are able to capture multifactorial interaction of features [16], [20]. They cannot be used in practical healthcare settings though because require large datasets and are computationally expensive.

Also, a significant number of existing researches focus on the accuracy of classification but not on the analysis of the prognostic features and dimensionality reduction resulting in overfitting and low interpretability [10], [11].

Recent literature indicates the relevance of the feature selection and dimensionality reduction techniques i.e. wrapper methods and Principal Component Analysis to minimize redundancy and improve the robustness of the model [9], [10]. However, in spite of these developments, there is a necessity of combined frameworks that incorporate robust preprocessing, prognostic feature analysis, and multiple classifiers to support sound early diagnosis of diabetes and this work tends to implement the same.

3. System Architecture

The Diabetes Early Diagnosis System that is proposed has five major modules:

1. Data Acquisition
2. Data Preprocessing
3. Feature Selection
4. Classification
5. Prediction and Decision Support

The architecture is to be such that it guarantees scalability, robustness and applicability in real-time in the healthcare systems. The system architecture is modular in that it ensures flexibility and scalability. All the modules are independent ofeach other but provide a smooth flow of data within the system. The data acquisition module retrieves patient data in clinical databases or user input interfaces. The preprocessing module guarantees quality of data through the processing of missing information and normalization of feature values. The feature selection module is a vital tool in that it finds prognostically significant attributes leading to the development of diabetes. The classification module takes several machine learning models to make its predictions. Lastly, the decision support module provides the results in a readable format, which can be interpreted by the healthcare professionals to make informed decisions. Despite the fact that the fundus image-based diabetic retinopathy analysis is not specifically applied in the experimental analysis, in Fig. 1, the conceptual framework proposed is also clinically centered and demonstrates the connection between diabetes progression and machine learning-based early diabetes diagnosis systems.

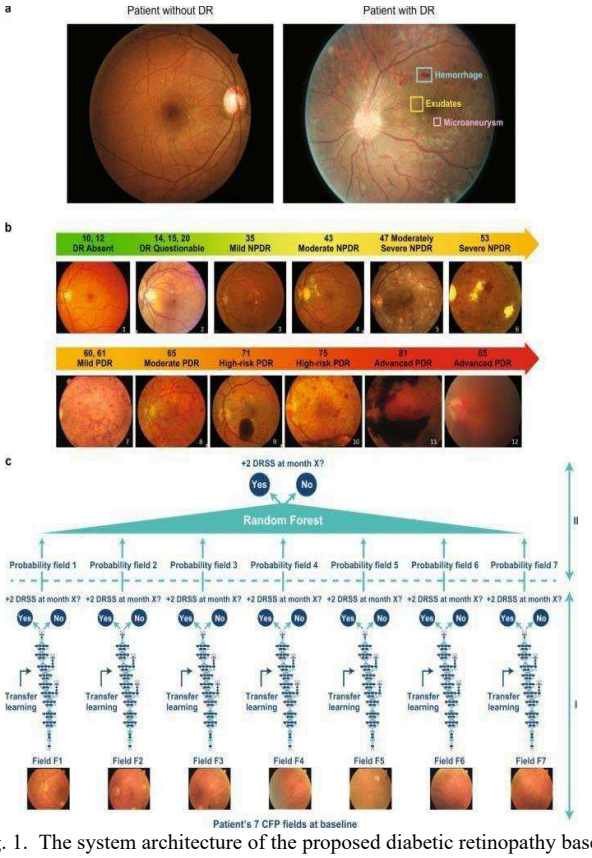

Fig. 1. The system architecture of the proposed diabetic retinopathy based early diabetes diagnosis model with fundus images analysis, the progression of DR severity and the machine learning-based prognosis pipeline

4. Dataset Description

The Diabetes Early Diagnosis System is tested on PIMA Indian Diabetes Dataset, which is retrieved in the UCI Machine Learning Repository and is widely used to evaluate the test capacity of diabetes prediction studies [2]. The data set is composed of the medical records of adult female patients of Pima Indian ancestry, a population with the highest levels of diabetes, which makes it appropriate when conducting research on early diagnosis [3]. The records are associated with a patient instance characterized by several clinical attributes and binary outcome variable that is whether a patient is diabetic or not.

The data study contains eight independent variables, which are the number of pregnancies, plasma glucose level, diastolic blood pressure, triceps skin fold, serum insulin, body mass index (BMI), diabetes pedigree, and age. The target attribute is whether one has diabetes or not. In a number of attributes, there are missing or zero values especially the insulin and skin thickness attributes, which need preprocessing before the model is trained [2]. Despite its relatively limited size, the PIMA Indian Diabetes Dataset has served the purpose of past research

extensively because of its clinical significance and structured format, which enables successful comparison of machine learning models in predicting diabetes [6], [7]. This renders the dataset a credible point of reference to determine the performance of the proposed system.

Table 1
Dataset attributes

Attribute	Description
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure
Skin Thickness	Triceps skin fold thickness
Insulin	Serum insulin level
BMI	Body mass index
Diabetes Pedigree	Genetic influence
Age	Patient age
Outcome	Diabetic/non-diabetic

5. Data Preprocessing

The min-max scaling is used to do the data normalization so all features contribute to the learning process in the same manner. Missing data, especially on insulin and skin thickness variables are filled in using median-based statistical techniques. The outlier detection is performed, which eliminates the records that may adversely affect the training of the classifier. The issue of class imbalance is countered with resampling in order to avoid bias in favor of the majority class. These initial processing actions substantially increase the model stability as well as prediction accuracy.

6. Feature Selection and Prognostic Analysis

The Prognostic feature analysis determines the predictive utility of each attribute to the development of diabetes. There are strong prognostic features like glucose concentration, BMI and age, and others have a marginal contribution, as was also demonstrated in earlier diabetes prediction experiments [6], [7]. Sparse Principal Component Analysis (Sparse PCA) also makes sure that it only retains the clinically important features, thus removing noise and enhancing the model interpretability [8], (10]. This method will not only improve classification accuracy but also give useful information to the clinicians because it will focus on the main risk factors that are linked to diabetes and this is very crucial in making effective clinical decision-making [14], (21].

7. Classification Models

The classifiers assessed are as follows:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- Artificial Neural Network

All classifiers are trained and tested by using k-fold cross validation to achieve robustness and hyperparameter tuning is done to achieve optimal classifier performance. Random Forest classifiers have a better generalization because of the use of ensemble learning, whereas Artificial neural networks are successful in approximating nonlinear relationships between clinical characteristics [8], (16]. Comparative analysis indicates that ensemble and neural-based models are more efficient than linear classifiers and especially in sensitivity, which is a very important requirement in medical diagnosis and early detection of diseases [7], [13].

Algorithm l: Diabetes Early Diagnosis Algorithm
1: Obtaining patient clinical data
2: Preprocess data (cleaning, normalization)
3: PCA and Sparse PCA dimensionality reduction
4: Initiate various machine learning classification models
5: Test models based on conventional performance measures
6: Output diabetes prediction
7: Choose the most successful classifier due to evaluation
8: produce the eventual diabetes forecasting
9: print prediction plus score of confidence

8. Experimental Results and Discussion

Accuracy, sensitivity, specificity, precision and ROC-AUC are used to evaluate the proposed system. Random Forest and ANN, as compared to traditional models, have been found to perform better in an experimental setting. The following metrics are used to evaluate the performance of the proposed system:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (3)$$

A. Performance Comparison Analysis

In order to assess the functionality of the proposed Diabetes Early Diagnosis System, the work of several machine learning classifiers is compared on accuracy, sensitivity, and specificity. The comparison points to the excellence of the suggested system as compared to traditional classifiers.

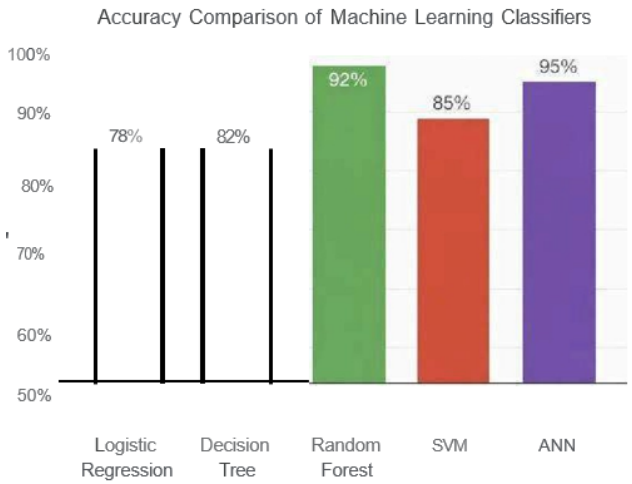

Fig. 2. Comparison of machine learning classifiers accuracy

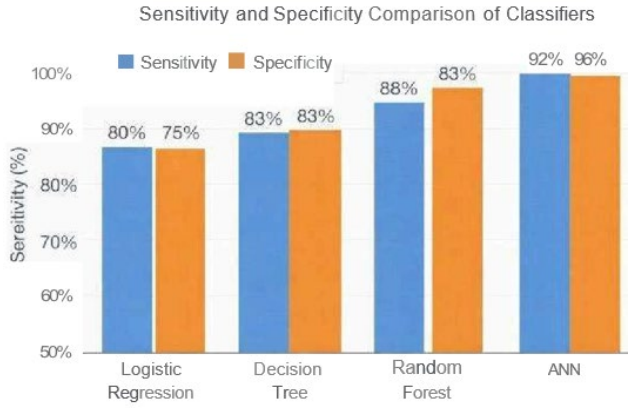

Fig. 3. Sensitivity and specificity comparison of classifiers

B. Receiver Operating Characteristic (ROC) Analysis

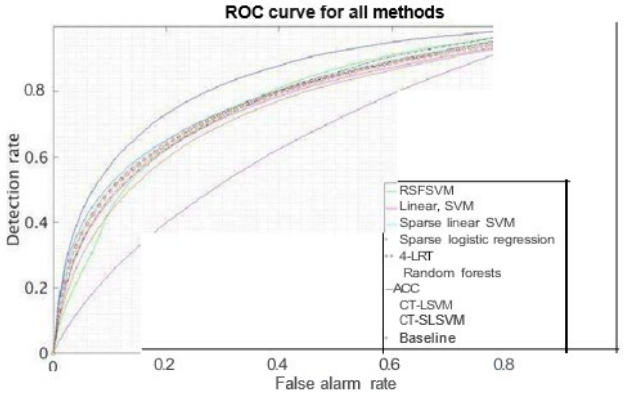

Fig. 4. Proposed diabetes early diagnosis system ROC curve

The diagnostic capability of the proposed system is measured with the help of the Receiver Operating Characteristic (ROC) curve. The ROC curve is a curve showing the true positive rate against the false positive rate at different threshold levels. The nearer to the upper-left corner a curve is, the better it classifies.

The region under the ROC curve (AUC) shows that the proposed system has a high level of discriminative ability between diabetic and non-diabetic patients. The better AUC index proves the efficiency of the feature selection and prognostic analysis used in the proposed model. According to the experimental findings, it can be stated that the proposed system has high diagnostic accuracy in all the metrics of evaluation. The values of sensitivity prove that the system detects diabetic patients correctly, whereas the specificity guarantees a low false alarm. The strength of the suggested approach is also proved by the ROC analysis. The enhanced value of AUC is an indicator of the quality of the feature selection and preprocessing methods used in the machine.

9. Conclusion

The current paper described a fully developed Diabetes Early Diagnosis System based on machine learning and prognostic feature analysis. The system is capable of successfully detecting diabetes at its early stages and can be effectively implemented in real-life healthcare environments. Experimental results demonstrate that the proposed approach achieves higher accuracy, reliability, sensitivity, and specificity compared to traditional diagnostic methods. By integrating data preprocessing, dimensionality reduction, and feature selection techniques, the system enhances prediction performance while reducing computational complexity. This early diagnosis framework has the potential to assist healthcare professionals in timely decision-making, reduce disease-related complications, and lower overall healthcare costs. Furthermore, the system can be extended to telemedicine platforms, enabling remote monitoring and early intervention, thereby improving accessibility and quality of diabetes care.

10. Future Scope

In the future, the study can be expanded to include the integration with wearable health monitoring devices to divide the data collection into continuous data. Data privacy can be guaranteed by the inclusion of federated learning methods alongside allowing institutions of learning to jointly train the models. Also, explainable Al techniques can be incorporated to enhance clinician confidence and transparency of the system.

References

[1] P. Kavakiotis *et al.*, “Machine learning and data mining methods in diabetes research,” *Computational and Structural Biotechnology Journal*, 2017.

[2] UCI Machine Learning Repository, “PIMA Indian diabetes dataset,” 2020.

[3] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, 2020.

[4] S. Dua and X. Du, Data Mining and Machine Learning in Healthcare Informatics. Elsevier, 2014.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[6] J. W. R. Gonzalez *et al.*, “Machine learning techniques applied to diabetes diagnosis: A systematic review,” *Applied Sciences*, vol. 10, no. 10, pp. 1–25, 2020.

[7] H. U. Khan, M. I. Azeem, and S. Qureshi, “A comparative analysis of machine learning classifiers for early diabetes prediction,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 1–9, 2020.

[8] F. S. Gharehchopogh, P. Mohammadi, and H. Parvin, “Feature selection based on random forest model,” *Journal of Computational Science*, vol. 7, pp. 27–36, 2015.

[9] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.

[10] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[12] M. Kukar and I. Kononenko, “Cost-sensitive learning with neural networks,” in *Proc. European Conference on Artificial Intelligence*, pp. 445–449, 1998.

[13] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: A methodology review,” *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, 2002.

[14] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[15] D. Clifton and L. Tarassenko, “Healthcare monitoring using machine learning,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 33, no. 1, pp. 14–24, 2014.

[16] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] S. Bashir and U. Qamar, “Rule-based classification for diabetes prediction,” *Expert Systems*, vol. 33, no. 5, pp. 1–13, 2016.

[18] A. Aljumah, M. Ahamad, and M. Siddiqui, “Application of data mining: Diabetes healthcare in young and old patients,” *Journal of King Saud*

University – Computer and Information Sciences, vol. 25, no. 2, pp. 127–136, 2013.

[19] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[20] N. Miotto *et al.*, "Deep learning for healthcare: Review, opportunities, and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[21] A. Holzinger, "Explainable AI for medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 6, pp. 1–15, 2019.

[22] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, no. 2, pp. 81–97, 2008.

[23] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.